Some Geometric Approaches to Parameter Estimation

David Hitchcock

Submitted for the degree of PhD

University of Edinburgh, 1992

This thesis has been composed by myself.

CONTENTS

ACKNOWLEDGEMENTS

I acknowledge the SERC for the necessary financial assistance for assuring the completion of this thesis.

I would like to thank my supervisor Terry Lyons for suggesting research topics and for helpful discussions.

I also thank Michael Rockner for his interest while he was in Edinburgh.

The following are due varying amounts of gratitude:

John Lunt for interesting conversations, occasionally on mathematics;

Mathematica and the late EMAS for some of the less interesting calculations;

Winston Sweatman for helping to search for enlightenment in some of Scotlands deepest bogs;

Ingrid, Susan and Hugh for helping to fight dragons;

Caledon Park Harriers for Athletic Support.

ABSTRACT

The major part of this thesis is concerned with some geometric aspects of a parametric statistical problem.

In chapter 1 we show how to assign structure to the parameter space by turning it into a Riemannian manifold. This is achieved by obtaining a metric from the model and the observations in a natural way. Different choices exist for the metric, and it is shown in later chapters that for specific problems this choice can be important. We also show how the standard idea of interest and nuisance parameters fits into this context.

In chapter 2 the gradient log-likelihood vector field is introduced and a natural diffusion process is put on the parameter space with this vector field as drift. Some properties of this diffusion are investigated including the relationship with the original statistical problem. A method for creating a diffusion on the interest parameter space by projection is introduced.

Chapter 3 considers the case when the nuisance parameters are incidental, which means that the the number of parameters increases with number of observations. In cases where an optimum method for dealing with the incidental parameters exists, it is shown that the method of chapter 2 is equivalent for the right choice of metric. The method is also applied to some more general cases and some of the problems that arise are explored.

Chapter 4 consists mainly of examples including a dynamic algorithm for computing properties of the likelihood measure which can be helpful in practice when other methods are intractable. The mixture model example is probably the most interesting.

Chapter 5 is somewhat disconnected from chapters 1-4 and considers observing a parameter dependent (continuous time) stochastic process at discrete time points. The actual likelihood function crucial to the analysis in chapters 1-4 cannot be calculated explicitly and so an alternative approach based on martingale techniques is presented.

Chapter 6 is almost entirely self-contained and presents a theorem on the detection of a signal when corrupted by white noise. The technique used is to consider the arrival point, θ as a parameter with a prior distribution U(0,1). Then observe the process at discrete times $\{k2^{-n}:k=0,1,\ldots,2^n\}$ (as in chapter 5) and thus obtain a log-likelihood function $l_n(\theta)$. Standard convergence theorems give that the posterior distribution for θ, $\exp[l_n(\theta)]/\int\exp[l_n(\theta)]d\theta$ converges weakly to the posterior distribution for θ given the whole path. Then properties of $l_n(\theta)$ as a stationary gaussian process are used to decide whether the posterior distribution for θ given the whole path can be a delta function at the true value or not.

This approach has not been used in previous papers on the problem, and while the proof is messy in places it does lead to a sharp result.

Chapter 1. The Geometry of a parametric statistical problem

The Model

The first part of this chapter shows how geometric considerations are a natural aspect of a standard parametric statistical model. A parametric model is specified by giving a set of probability distributions for some random variable X indexed by a parameter θ. The idea is that one of these distributions is in fact true and X is a random variable from that distribution. A particular observation x will be more likely under some of the parameters than others, and the problem is to infer what we can about the true parameter value. It is usually the case that all the distributions are absolutely continuous, and so observation of a particular x will not allow us to rule out any of the parameters.

The geometric aspect stems from the assumption that the set of possible distributions takes the form of an m-dimensional smooth manifold, M. A manifold is a set which locally looks like $\mathbb{R}^m$, though there may be more global structure. The definition is in terms of an atlas of homeomorphisms φ_θ between subsets of M and subsets of $\mathbb{R}^m$. For every point θ in the manifold there is a neighbourhood U_θ of θ and a map $\varphi_\theta : U_\theta \longrightarrow \mathbb{R}^m$ with $\varphi_\theta(\theta) = 0$ which is a homeomorphism onto its range. The smoothness comes from a consistency requirement which is that if $U_\theta \cap U_{\theta'} \neq \emptyset$ then $\varphi_\theta \circ \varphi_{\theta'}^{-1} : \varphi_{\theta'}(U_\theta \cap U_{\theta'}) \longrightarrow \varphi_\theta(U_\theta \cap U_{\theta'})$ is not only a homeomorphism but also smooth (C^∞).

In typical problems the possible distributions are indexed by real numbers and so the space of possible distributions will automatically be a manifold, henceforth called the parameter space.

Once we have the parameter space, the model can be fully defined by specifying a map $X: M \times \Omega \longrightarrow \mathfrak{X}$ where Ω is a probability space and $\mathfrak{X}$ is the sample space. This map is smooth in the first argument and measurable in the second. The quantity X, which represents what can be observed from performing the experiment, is called a random variable.

Problems are not normally given in terms of such a map; a 1-dimensional model might be specified by giving the distribution function $F_\theta(.)$ of a real-valued random variable, X, so that $F_\theta(x) = \mathbb{P}[X \leqslant x]$. This can be formulated by taking $\mathfrak{X} = \mathbb{R}$ and $\Omega = ([0,1], Leb)$ and defining $X(\theta,\omega) = F_\theta^{-1}(\omega)$.

For each $\theta \in M$ we have $X|_\theta: \Omega \longrightarrow \mathfrak{X}$ and so we can derive a probability measure $\mathbb{P}_\theta$ on $\mathfrak{X}$: if $A \subset \mathfrak{X}$, $\mathbb{P}_\theta(A) = meas\{X|_\theta^{-1}(A)\}$. We assume that these measures are all absolutely continuous with respect to some measure dx on $\mathfrak{X}$, so we have a density function $f(x;\theta)$. While different versions may exist for the density, only one can be continuous; we assume that for each θ the density is a smooth function $\mathfrak{X} \longrightarrow \mathbb{R}^+$. This function is called the likelihood when considered as a (random) function $M \longrightarrow \mathbb{R}^+$ and the log-likelihood $l(\theta) = \log[f(x;\theta)]$. This is a smooth function $M \longrightarrow \mathbb{R}$ and is only defined up to an additive constant because if a different reference measure had been chosen on $\mathfrak{X}$ then the density function would be $f(x;\theta)\psi(x)$ for some ψ.

We suppose that one of the parameter values, θ_0 is in fact true but unknown and $X(\theta_0,\omega) = x$ is observed. All values of $\theta \in M$ remain possible but in typical cases, if certain values of θ were true x would be a very unlikely observation while if other values of θ were true x would be a reasonable observation. Our problem is to quantify belief in the different parameter values.

A standard approach is to find the maximum likelihood estimator, $\hat{\theta}(x)$ satisfying $l(\hat{\theta}) \geq l(\theta)\ \forall\ \theta \in M$. The standard results concerning mle's come from the case of iid sampling, when we have a sequence $\{X_i\}$ of independent and identically-distributed random variables. Let $\hat{\theta}(\kappa) \in M$ be the mle based on the observations $X_1,\ldots,X_\kappa$. Under fairly mild regularity conditions the sequence $\hat{\theta}(\kappa)$ converges to the θ_0 (consistency) and $\sqrt{\kappa}(\hat{\theta}(\kappa)-\theta_0)$ has an asymptotically normal distribution - see Cox[4] p294. However, examples will be considered in this thesis where the mle is not consistent. In any case a point estimate is of limited value on its own as it is almost surely incorrect. A sequence of point estimates may be useful because of the above asymptotic results, but to be strictly in accordance with the model defined there is only one observation and so only one mle, as the size of the experiment has been fixed in advance. Of course this observation may be a vector of fixed length.

If the parameter were selected from a known prior distribution, then Bayes Rule gives a posterior distribution from which meaningful statements about the relative probabilities of different regions of M can be made. If no genuine prior exists we could use an improper prior and proceed in the same way; the

improper posterior should not be interpreted in the same direct way, but a likelihood measure on the parameter space can be a useful tool. This is partly because in many situations this likelihood measure is fairly robust with respect to changing the improper prior and can therefore be reasonably interpreted as a posterior measure. A different method for producing a likelihood measure on the interest parameter space is presented here which can perform as well, and better in some cases.

Hypothesis testing methods (including confidence regions) generally give useful results though the technique may be hard to implement when there are nuisance parameters. Their pre-experimental nature can lead to perverse conclusions: see Berger[2] p5.

Metrics

The direction we take is to use the model to put more structure on the parameter space. At the moment M is simply the set of possible distributions.

Example 1.1: Let $\{X_i : i \leqslant \kappa\}$ be an iid sample from $N(\mu,\sigma^2)$. The parameter space would naturally be $\mathbb{R}\times\mathbb{R}^+$. However we could choose different coordinates and make $\zeta=(1+\sigma^2 e^{\mu})^{-1}$ and $\xi=(1+\sigma^2)^{-1}$. Any of the possible distributions is represented by some (ζ,ξ) in the unit square, so the parameter space could equally well be the unit square.

■

We can fix one particular representation of the parameter space by making it a Riemannian Manifold. This is done by defining an inner product, $\langle .,. \rangle$, on the tangent space at each point (in a smooth way). This will fix the size, shape and curvature of the manifold. This inner product should not depend on the coordinate chart initially chosen. In coordinates θ, the standard basis vectors of the tangent space at a fixed point are written $\partial/\partial\theta_i$. FIGURE 1.1 shows a parameterization on a curved manifold where the basis tangent vectors at (2,2) are not orthogonal or of unit length. The inner product is exhibited as an m×m positive definite matrix, $g_{ij} = \langle \partial/\partial\theta_i, \partial/\partial\theta_j \rangle$, which will depend on position in the manifold. As $\langle .,. \rangle$ does not depend on any particular coordinate system, the matrix-valued function g must transform as a tensor.

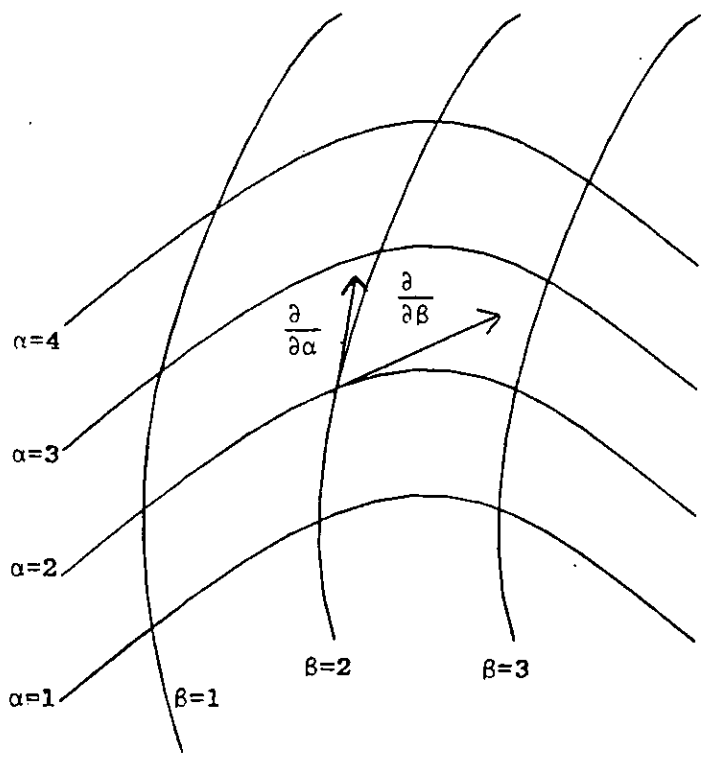


FIGURE 1.1

The volume element associated to the metric is given by $\left[\text{volume of the parallelpiped spanned by } \{\frac{\partial}{\partial\theta_1},\ldots,\frac{\partial}{\partial\theta_m}\}\right]d\theta_1\ldots d\theta_m = |g|^{1/2}d\theta_1\ldots d\theta_m$, where $|g| = \det(g)$.

The Fisher Information can be used as a metric - see Amari[1] p25. Let α and β be tangent vectors in TM_θ, the tangent space to M at θ. Since l is a (random) $\mathbb{R}$-valued function on M (0-form) we have that $\alpha l \in \mathbb{R}$. Define $g(\alpha,\beta) = \mathbb{E}[\alpha l.\beta l]$ or in coordinates $g_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial\theta_i}.\frac{\partial l}{\partial\theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right]$. The second equality follows from integration by parts (given suitable smoothness conditions) - see Cox[4] p108.

Example 1.2: Let $\{X_i : i \leqslant \kappa\}$ be an iid sample from $N(\mu,\sigma^2)$.

Then $l = -\frac{\kappa}{2}\log(\sigma^2) - \Sigma(x_i-\mu)^2/2\sigma^2$

$\partial l/\partial\mu = \Sigma(x_i-\mu)/\sigma^2$, $\partial l/\partial\sigma^2 = -\frac{\kappa}{2\sigma^2} + \Sigma(x_i-\mu)^2/2\sigma^4$

$\frac{\partial^2 l}{\partial\mu^2} = -\kappa/\sigma^2$, $\frac{\partial^2 l}{\partial\mu\partial\sigma^2} = -\Sigma(x_i-\mu)/\sigma^4$, $\frac{\partial^2 l}{(\partial\sigma^2)^2} = \frac{\kappa}{2\sigma^4} - \Sigma(x_i-\mu)^2/\sigma^6$.

So $g = \begin{bmatrix}\kappa/\sigma^2 & 0\\ 0 & \kappa/2\sigma^4\end{bmatrix}$

The volume element derived from the metric is given by $|g|^{1/2}d\mu d\sigma^2 = \kappa/\sigma^3\sqrt{2}\ d\mu d\sigma^2$

■

The following is motivation that the Fisher Information is the right metric to use - at least close to the true value: let $\mathcal{E}: U_{\theta_0} \longrightarrow TM_{\theta_0}$ be a projection from a neighbourhood of θ_0 to the tangent space at θ_0. The map $\mathcal{E}$ can be derived from one of the homeomorphisms used in the definition of the manifold. Recall we have $\varphi_{\theta_0}: U_{\theta_0} \longleftrightarrow \mathbb{R}^m(0)$ a homeomorphism between U_{θ_0} and a

neighbourhood of 0 in $\mathbb{R}^m$. Suppose the axes of $\mathbb{R}^m$ are labelled $\theta_1,\ldots,\theta_m$. Each coordinate function in $\mathbb{R}^m$ gives a basis vector, $\partial/\partial\theta_i$ in TM_{θ_0}. Then if $\varphi_{\theta_0}(\theta)=(\alpha_1,\ldots,\alpha_m)$ then $\mathcal{E}(\theta) = \sum_i \alpha_i \partial/\partial\theta_i$. The canonical example for $\mathcal{E}$ is the inverse of the exponential map.

Lemma: In the iid case when the mle is asymptotically normal, the limiting distribution of $\mathcal{E}(\hat{\theta})$ is N(0,I), the standard normal distribution. Note that the metric has size proportional to κ, the number of observations so the squared distance between two fixed points on the manifold will be proportional to κ.

Proof: Translate coordinates so that $\theta_0=0$. The standard result (see Cox[4] p294) is that the limiting distribution of $\hat{\theta}\sqrt{\kappa}$ is $N[0,i^{-1}(\theta_0)]$ where $i(\theta_0)$ is the Fisher Information for one observation. So the standard result is that $\Phi_{\theta_0}(\hat{\theta}\sqrt{\kappa}) \longrightarrow (\alpha_1,\ldots,\alpha_\kappa)^T$ in distribution, where $\mathbb{E}[\alpha_i] = 0$ and $\mathbb{E}[\alpha_i\alpha_j] = i^{-1}(\theta_0)_{ij}$.

Now $\mathcal{E}(\hat{\theta})\sqrt{\kappa} \longrightarrow \Sigma\alpha_i\ \partial/\partial\theta_i \in TM_{\theta_0}$, and

$\mathbb{E}[<\mathcal{E}(\hat{\theta}),\partial/\partial\theta_i><\mathcal{E}(\hat{\theta}),\partial/\partial\theta_j>] \longrightarrow \mathbb{E}[\alpha_r\alpha_q/\kappa\ g_{ri}g_{qj}] = g_{ij} = <\partial/\partial\theta_i,\partial/\partial\theta_j>$, where summation is implied by repeated suffices.

■

It can be more useful to use the observations directly to make the metric. If independent observations $x_1,\ldots,x_\kappa$ have log-likelihoods $l_1,\ldots,l_\kappa$ so that $l = l_1+\ldots+l_\kappa$ then the empirical metric is defined: $h(\alpha,\beta) = \sum_{r=1}^{k} \alpha l_r\ \beta l_r$ or in

coordinates, $h_{ij} = \frac{\partial l_r}{\partial \theta_i}\frac{\partial l_r}{\partial \theta_j}$. Under this metric, the manifold naturally embeds in $\mathbb{R}^\kappa$ as $\{(l_1(\theta),\ldots,l_\kappa(\theta)): \theta \in M\}$. If $\kappa < m$ the metric is singular everywhere, which has the statistical interpretation that if there are fewer observations than parameters then it will be impossible to distinguish between parameters. If $\kappa = m$ the metric is in general singular on a $\kappa-1$ dimensional submanifold, since we cannot expect the map from M to $\mathbb{R}^\kappa$ to be one-one.

An alternative metric is $\tilde{h}_{ij} = \frac{\partial(l_r-\bar{l})}{\partial\theta_i}.\frac{\partial(l_r-\bar{l})}{\partial\theta_j}$, where $\bar{l} = l/\kappa$ is the mean likelihood. The significance of subtracting the mean likelihood at each stage is that the vector $\sum_{r=1}^{t} \frac{\partial l_r}{\partial\theta_i}$ is a martingale; empirical estimation of its bracket is made by subtracting $\frac{\partial\bar{l}}{\partial\theta_i}$ at each stage so that the overall increment is 0. To visualize this metric project the embedding given for h above to the plane $x_1+\ldots+x_\kappa = 0$. If $\kappa \leqslant m$ this metric is singular everywhere; if $\kappa = m+1$ it is in general singular on a $\kappa-1$ dimensional submanifold, though not in the exponential family case (see below).

Example 1.9: Let $\{X_i : i \leqslant \kappa\}$ be an iid sample from $N(\mu,\sigma^2)$.

$$l_r = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x_r-\mu)^2. \qquad \frac{\partial l_r}{\partial\mu} = (x_r-\mu)/\sigma^2.$$

$$\frac{\partial l_r}{\partial\sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x_r-\mu)^2$$

$$g = \begin{bmatrix} \kappa/\sigma^2 & 0 \\ 0 & \kappa/2\sigma^4 \end{bmatrix}$$

$$h = \begin{bmatrix} \Sigma(x_r-\mu)^2/\sigma^4 & \Sigma(x_r-\mu)^3/2\sigma^6 - (x_r-\mu)/2\sigma^4 \\ \Sigma(x_r-\mu)^3/2\sigma^6 - (x_r-\mu)/2\sigma^4 & \Sigma(x_r-\mu)^4/4\sigma^8 - (x_r-\mu)^2/2\sigma^6 + 1/4\sigma^4 \end{bmatrix}$$

The metric $\tilde{h}$ in the exponential family case:

The metric $\tilde{h}$ is not best represented in these coordinates. The log-likelihood for the rth observation from any exponential family can be written $l_r = \Sigma\ \theta_j \tau_j^{(r)} - \psi(\theta)$ in terms of the natural parameter θ and the natural statistic τ. Then $l_r - \bar{l} = \theta_j\left[\tau_j^{(r)} - \overline{\tau_j}\right]$ and $\frac{\partial(l_r-\bar{l})}{\partial\theta_j} = \tau_j^{(r)} - \overline{\tau_j}$, where $\overline{\tau_j} = \sum_{r=1}^{K} \tau_j^{(r)}/\kappa$. So $\tilde{h}_{ij} = \sum_{r=1}^{K}\left[\tau_i^{(r)} - \overline{\tau_i}\right]\left[\tau_j^{(r)} - \overline{\tau_j}\right]$ which is independent of θ ie position on the manifold. Thus the vectors $\partial/\partial\theta_i$ and $\partial/\partial\theta_j$ have the same length and the same inner product everywhere on the Manifold. This proves that the parameter space of any exponential model is Euclidean Space under the metric $\tilde{h}$. The natural parameters are not necessarily the Euclidean coordinates. This is because the matrix $\tilde{h}_{ij}$ depends on the observations; If the $\{\theta\}$ were Euclidean we would need $\tilde{h}_{ij} \equiv \delta_{ij}$.

In the normal case we have $\theta_1 = 1/2\sigma^2$, $\theta_2 = \mu/\sigma^2$, $\tau_1^{(r)} = -x_r^2$, $\tau_2^{(r)} = x_r$, $\tilde{h} = \kappa\begin{bmatrix} \overline{x^4}-\overline{x^2}^2 & -\overline{x^3}+\overline{x^2}\,\overline{x} \\ -\overline{x^3}+\overline{x^2}\,\overline{x} & \overline{x^2}-\overline{x}.\overline{x} \end{bmatrix}$ ■

Example 1.4: Possible Disconnectedness

Let θ be an angle between 0 and π and let $f_X(x) = (2/\pi)\cos^2(x-\theta)$ be the density where $x \in [0,\pi)$. With two observations, we embed the parameter space under the metric h by plotting $(\log[\cos^2(x_1-\theta)], \log[\cos^2(x_2-\theta)])$. Typically this will have two disconnected branches. This highlights the need to consider the distance between two distributions as their distance apart in $\mathbb{R}^K$ and not their distance apart on the manifold; taking account of this is not always easy to implement.

For the case of the Fisher Metric,

$$\log f_X(x) = \log[\cos^2(x-\theta)]$$

$$\frac{\partial}{\partial x}\log f_X(x) = 2\tan(x-\theta)$$

$$\frac{\partial^2}{\partial x^2}\log f_X(x) = -2\sec^2(x-\theta)$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial x^2}\log f_X(x)\right] = \int_0^{\pi} (4/\pi)\sec^2(x-\theta)\cos^2(x-\theta) = 4$$

So under the Fisher Metric θ is a Euclidean coordinate and the parameter space is just a circle.

■

The parameter space being disconnected may cause problems when we put a diffusion process on M since the diffusion will be restricted to one of the branches. This example highlights a difference between the two metrics. To find the information at a point θ in the manifold the Fisher Metric assumes θ to be true and then gives the long-term expected information. The Empirical Metric assigns the information that the data suggest is available.

Interest Parameters

In many statistical problems, not all the parameters are of interest. In the geometrical context we have an interest parameter space N, which is a manifold of dimension less than that of M, and a natural map $\varphi: M \longrightarrow N$. This map is assumed to be a smooth submersion which means that if $\varphi(\theta) = \lambda$ then the derivative $\varphi'(\theta): TM_\theta \longrightarrow TM_\lambda$ has full rank $dim(N)$ $\forall \theta \in M$. A consequence of this is that for each $\lambda \in N$, $\varphi^{-1}(\lambda)$ is a submanifold of M, called the fibre above λ, and we have a foliation of M into such fibres (see Hicks [12] for details).

From the statistical standpoint we would like to use a model where the parameter space was just N. However it may prove impossible to satisfactorily model the experiment as such, so for each value of the interest parameter we have to consider a range of possible distributions so that the total parameter space is a larger manifold M.

The quantities such as the likelihood and the metric are all defined on M. In order to make inferences about λ we need to obtain functions on N. To be statistically significant, these must be obtained in a way which is independent of any coordinate system for the fibres $\varphi^{-1}(\lambda)$. This will be ensured by considering geometrical quantities so that procedures can at least be formulated in a coordinate-free way.

Chapter 2 Diffusion Processes

The metric should be thought of as using the model and maybe secondary information from the observations to give structure to the parameter space. A metric fixes distance and orthogonality on the manifold, and the distance apart of two parameter values represents how easily they can be distinguished.

The primary information from the observations arises from using ∇l: this is a vector field pointing in the direction of most rapidly-increasing log-likelihood. The concept of 'most rapidly increasing' requires a concept of distance which requires a metric. A particle moving along this vector field, ie always with velocity ∇l will converge to the maximum likelihood estimate(or possibly only a local maximum). However the mle cannot really be distinguished from its close neighbours as possible parameter values: this is incorporated by adding to the path of the particle a random term to blur the distinction between two points which are close on the manifold.

Following Rogers and Williams[19] V-30 we can construct a diffusion process $\Re$ on M with generator $\mathcal{G} = \frac{1}{2}\Delta + \frac{1}{2}\nabla l.\nabla$ where ∇ is gradient, $\nabla.$ is divergence and $\Delta = \nabla.\nabla$ is the Laplace-Beltrami Operator, all of which depend on the metric. This diffusion is Brownian motion on the manifold with drift $\frac{1}{2}\nabla l$ and can be symbolically written $d\Re = dW + \frac{1}{2}\nabla l\ dt$.

Stationary distribution

The generator of the process $\mathcal{G} = \frac{1}{2}\Delta + \frac{1}{2}\nabla l.\nabla$ is related to the process itself by the Hille-Yosida Theorem. If $f:M \longrightarrow \mathbb{R}$ is C^2 then:

$$\mathbb{E}[f(\aleph_{t+\varepsilon})|\aleph_t] = f(\aleph_t) + \varepsilon\mathcal{G}f(\aleph_t) + o(\varepsilon)^{[\dagger]}.$$

Now if $\aleph_t$ has the stationary distribution then for any $A \subset M$ we must have:

$$\mathbb{P}[\aleph_t \in A] = \mathbb{P}[\aleph_{t+\varepsilon} \in A] \text{ and therefore}$$

$$\mathbb{E}[f(\aleph_{t+\varepsilon})] = \mathbb{E}f(\aleph_t) \text{ holds for all measureable } f.$$

So taking expectations of $^{[\dagger]}$ we must have that

$$0 = \mathbb{E}[\mathcal{G}f(x)] = \int \mathcal{G}f(x)\ \mu(x)dx$$

where μ is the stationary measure. So

$$0 = \int f(x)\ \mathcal{G}^*\mu(x)dx \text{ where } \mathcal{G}^* \text{ is the adjoint of } \mathcal{G}.$$

Since this holds for all $f \in C^2$ the stationary measure is given by the minimal solution to the Kolmogorov Forward Equation $\mathcal{G}^*\mu = \frac{1}{2}\Delta\mu - \frac{1}{2}\nabla.(\mu\nabla l) = 0$.

Setting $\mu = e^l$ we have that $\mathcal{G}^*e^l = \frac{1}{2}\nabla.(\nabla e^l - e^l\nabla l) = 0$ so the stationary distribution is simply the likelihood with respect to the volume element of the metric (normalized to be a probability distribution). This shall be called the likelihood measure.

This describes the diffusion intrinsically on the parameter space, ie independently of any particular coordinate system. We will need to look at coordinates when there are interest parameters and also to simulate the diffusion. So suppose that we have a coordinate system θ and can view the

parameter space (locally) as $\mathbb{R}^m$. The metric now takes the form of a positive definite m×m matrix g at each point.

$$\mathcal{G} = \tfrac{1}{2} \quad \nabla_{\cdot} \quad \nabla \quad + \tfrac{1}{2} \quad \nabla l \quad \cdot \quad \nabla$$

$$= \tfrac{1}{2} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial\theta_i} \left[\sqrt{|g|} \quad g^{-1}_{ij} \frac{\partial}{\partial\theta_j}\right] + \tfrac{1}{2} g^{-1}_{ij} \frac{\partial l}{\partial\theta_j} \quad g_{ih} \quad g^{-1}_{hk} \frac{\partial}{\partial\theta_k}$$

$$= \frac{1}{2} g^{-1}_{ij} \frac{\partial^2}{\partial\theta_i\partial\theta_j} + \frac{1}{2}\left[\frac{\partial l}{\partial\theta_i} g^{-1}_{ij} + \frac{1}{\sqrt{|g|}} \frac{\partial\sqrt{|g|}}{\partial\theta_i} g^{-1}_{ij} + \frac{\partial}{\partial\theta_i}[g^{-1}_{ij}]\right]\frac{\partial}{\partial\theta_j}$$

$$\mathcal{G}^*\mu = \frac{1}{2} \frac{\partial^2}{\partial\theta_i\partial\theta_j}\left[\mu\, g^{-1}_{ij}\right] - \frac{1}{2} \frac{\partial}{\partial\theta_j}\left[\mu\left[\frac{\partial l}{\partial\theta_i} g^{-1}_{ij} + \frac{1}{\sqrt{|g|}} \frac{\partial\sqrt{|g|}}{\partial\theta_i} g^{-1}_{ij} + \frac{\partial}{\partial\theta_i}[g^{-1}_{ij}]\right]\right]$$

This is solved by $\mu = e^{l}\sqrt{|g|}$ as $\sqrt{|g|}$ is the volume element according to the metric g.

To write the diffusion in coordinates, we need a symmetric matrix whose square is g^{-1}.

The diffusion is then:

$$d\theta_j(t) = g^{-1/2}_{ij} dB_i(t) + \frac{1}{2}\left[\frac{\partial l}{\partial\theta_i} g^{-1}_{ij} + \frac{1}{\sqrt{|g|}} \frac{\partial\sqrt{|g|}}{\partial\theta_i} g^{-1}_{ij} + \frac{\partial}{\partial\theta_i}[g^{-1}_{ij}]\right]dt$$

where B(t) is an $\mathbb{R}^m$ Brownian motion. For the metric h [resp. $\tilde{h}$], the stochastic part can be written:

$$h^{-1/2}_{ij} dB_i(t) = h^{-1}_{ij} \frac{\partial l_r}{\partial\theta_i} d\tilde{B}_r(t)$$

$$[\text{resp. } \tilde{h}^{-1/2}_{ij} dB_i(t) = \tilde{h}^{-1}_{ij} \frac{\partial[l_r - \bar{l}]}{\partial\theta_i} d\tilde{B}_r(t)\ ]$$

for some $\mathbb{R}^K$ Brownian motion $\tilde{B}$. The proof that both formulations have the same distribution is a simple application of Levy's Theorem.

Example 2.1: For the case of a normal distribution with unknown mean and variance we have

$$l = -\frac{\kappa}{2}\left[\log(\sigma^2) + \frac{\overline{x^2}}{\sigma^2} - \frac{2\mu\bar{x}}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right].$$

$$= -\frac{\kappa}{2}\left[\frac{1}{\sigma^2}(\mu-\bar{x})^2 + \frac{s_2}{\sigma^2} + \log(\sigma^2)\right], \text{ where } s_2=\overline{x^2}-\bar{x}^2.$$

$$g = \begin{bmatrix}\kappa/\sigma^2 & 0\\ 0 & \kappa/2\sigma^4\end{bmatrix}$$

The volume element for the Fisher Metric is $\sqrt{|g|}d\mu d\sigma^2 = \frac{\kappa}{\sigma^3\sqrt{2}}\, d\mu d\sigma^2$. So the stationary distribution of the diffusion has measure proportional to

$$e^{-\kappa(\mu-\bar{x})^2/2\sigma^2}(\sigma^2)^{-\kappa/2-3/2}e^{-\kappa s_2/2\sigma^2}\, d\mu d\sigma^2$$

$$= \sigma^{-1}e^{-\kappa(\mu-\bar{x})^2/2\sigma^2}(\theta)^{(\kappa+2)/2}e^{-\kappa s_2\theta/2}\,\theta^{-2}d\mu d\theta \text{ where } \theta = 1/\sigma^2$$

Hence the stationary distribution has

$$\mu|\sigma^2 \sim N(\bar{x},\sigma^2/\kappa), \quad 1/\sigma^2 \sim \Gamma(\kappa s_2/2,\kappa/2)$$

The diffusion itself can be written:

$$d\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} = \begin{bmatrix}\sigma & 0\\ 0 & \sigma^2\sqrt{2}\end{bmatrix}\begin{bmatrix}dB_1\\ dB_2\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\begin{bmatrix}\Sigma(x_i-\mu)^2/2\sigma^2\\ -\kappa/2\sigma^2+\Sigma(x_i-\mu)^2/2\sigma^4\end{bmatrix}dt +$$

$$\frac{\sigma^3\sqrt{2}}{2}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\begin{bmatrix}0\\ -2/2\sigma^5\sqrt{2}\end{bmatrix}dt + \frac{1}{2}\begin{bmatrix}0\\ 4\sigma^6\end{bmatrix}dt$$

$$= \begin{bmatrix}\sigma dB_1 + \tfrac{1}{2}\Sigma(x_i-\mu)^2dt\\ \sigma^2\sqrt{2}dB_2 + \tfrac{1}{2}\{\Sigma(x_i-\mu)^2-\kappa\sigma^2+\sigma^2\}dt\end{bmatrix}$$

Under the current set-up, a more natural diffusion process, though perhaps more complicated to analyse, is the diffusion Ω on the tangent bundle (TM) corresponding to Ornstein-Uhlenbeck process with drift. This has two components: θ is position on the manifold and $V \in TM_\theta$ is velocity. We want to define a process where V is affected by drag, drift and noise and so the equations can be symbolically written:

$$d\theta = Vdt;\ dV = (-2cV + c\nabla l)dt + 2cdW$$

where W is a Brownian Motion $c \in \mathbb{R}^+$ is a constant. Such an equation does not make complete sense as V lives in a different space at each time.

Given a path in M from θ to θ' and a vector in TM_θ , as a particle moves along the path in M so the vector can simultaneously be mapped to the tangent space above the particle by means of a connexion on the manifold, so-called because it connects neighbouring tangent spaces. Amari[1] considers a 1-parameter family of connexions, but the only natural choice here is the one derived directly from the metric - the Levi-Civita connexion (Amari's 0-connexion) - as this is the only connexion which preserves the inner product (ie the energy). A particle starting at $\theta \in M$ with initial velocity $v \in TM_\theta$ with no additional forces acting will move over the manifold at the same speed, in fact following a geodesic. The diffusion will now be rigorously defined in terms of its generator. When defined in a coordinate-free way, the role of the Levi-Civita connexion is implicit

The generator of this process acts on C^2 functions f from the tangent bundle to the reals, $f:TM \longrightarrow \mathbb{R}$. Suppose we have $(\theta,v) \in TM$. The gradient in a tangent space, ∇_V is easy to define as for θ fixed, f maps an inner product space $(TM_\theta,<.,.>)$ to $\mathbb{R}$. Gradient on the manifold, ∇, is less straightforward: recall we have U_θ, a neighbourhood of θ and $\varphi_\theta:U_\theta \longrightarrow \mathbb{R}^m(0)$ a homeomorphism to a neighbourhood of $0 \in \mathbb{R}^m$. Then for each $\theta' \in U_\theta$ we have $\varphi_\theta^{-1}(0,\varphi_\theta(\theta'))$ is a line joining θ to θ' and so we can move v

along this line using the Levi-Civita connexion to get $v' \in TM_{\theta'}$. Then we can define $f_v^{\varphi}: U_\theta \longrightarrow \mathbb{R}$ by $f_v^{\varphi}(\theta') = f(\theta', v')$ and define ∇f at (θ, v) by ∇f_v^{φ} at θ. This definition is independent of φ_θ in that if $\tilde{\varphi}_\theta$ were a different homeomorphism with $\tilde{\varphi}_\theta \circ \varphi_\theta^{-1}$ smooth, using $\tilde{\varphi}_\theta$ would lead to the same definition of ∇.

Following the formal definition above we look at the diffusion with the following quantity as generator:

$$\mathcal{G} = 2c^2 \nabla_V^2 + V.\nabla + (c\nabla l - 2cV).\nabla_V.$$

$$\mathcal{G}^* \mu = 2c^2 \nabla_V^2 \mu - \nabla.[V\mu] - \nabla_V.[\mu(c\nabla l - 2cV)].$$

Putting $\mu = e^{l} e^{-<v,v>/2c}$,

$$\mathcal{G}^* \mu = 2c^2(<v,v>/c^2 - m/c)\mu - \mu v.\nabla l + (v/c).(c\nabla l - 2cv)\mu + 2cm\mu$$

$$= 0.$$

So the stationary distribution has θ distributed as the likelihood with respect to the volume element of the metric g, and V independent and normal (0,c).

In local coordinates: $d\theta_i = V_i dt$

$$dV_i = \left[-2cV_i + cg_{ij}^{-1} \frac{\partial l}{\partial \theta_j} - V_j V_k \Gamma^i_{jk}\right] dt + 2cg_{ij}^{-1/2}\, dB_j.$$

where $\Gamma^i_{jk} = \frac{1}{2}\, g^{-1}_{ir} \left[\frac{\partial g_{rj}}{\partial \theta_k} + \frac{\partial g_{rk}}{\partial \theta_j} - \frac{\partial g_{jk}}{\partial \theta_r}\right]$ is the Christoffel function for the Levi-Civita connexion.

$$\mathcal{G} = 2c^2 g_{ij}^{-1} \frac{\partial^2}{\partial v_i \partial v_j} - 2cv_i \frac{\partial}{\partial v_i} + v_i \frac{\partial}{\partial \theta_i} + cg_{ij}^{-1} \frac{\partial l}{\partial \theta_j} \frac{\partial}{\partial v_i}$$

$$- v_j v_k \Gamma^i_{jk} \frac{\partial}{\partial v_i}$$

$$\mathcal{G}^* \mu = 2c^2 \frac{\partial^2}{\partial v_i \partial v_j}\left[\mu g_{ij}^{-1}\right] + 2c \frac{\partial}{\partial v_i}[\mu v_i] - \frac{\partial}{\partial \theta_i}[\mu v_i]$$

$$- c \frac{\partial}{\partial v_i}\left[\mu g_{ij}^{-1} \frac{\partial l}{\partial \theta_j}\right] + \frac{\partial}{\partial v_i}\left[\mu v_j v_k \Gamma^i_{jk}\right]$$

With $\mu = \sqrt{|g|}e^{l}\ \sqrt{|g|}e^{-v_i g_{ij} v_j/2c} = |g|e^{l}\ e^{-v_i g_{ij} v_j/2c}$ we have $\mathcal{G}^*\mu = 0$ so the stationary distribution is μ distributed as the likelihood measure and V independent and Normal(0,c) in the tangent space. This can be checked directly by using the formula for the Levi-Civita Connexion:

Example 2.2 - Saw tooth example: This is an example where the likelihood function is not smooth. Under the empirical metric $\tilde{h}$, the diffusion $\mathfrak{R}$ is awkward to define, but the same problems do not occur with the diffusion Ω. Let s(x) be the saw-tooth function, $s(x) = |x| : |x| \leq 1$, s periodic with period 2. Let $\theta \in (-1,1]$ be the parameter and $f(x;\theta) = e^{s(x-\theta)}/2(e-1) : |x| \leq 1$ be the density. Now for two observations to obtain the embedding under the metric h, the two log-likelihoods are plotted against each other. There will be four segments: one segment will correspond to both log-likelihoods increasing and one to both decreasing, and these will give straight line segments in $\mathbb{R}^2$ perpendicular to the line $l_1+l_2=0$; and two segments correspond to one log-likelihood increasing and one decreasing, and these will give straight line segments in $\mathbb{R}^2$ parallel to the line $l_1+l_2=0$. Thus the parameter space embeds as a rectangle. When projected down to the line $l_1+l_2=0$, the former two sides get squashed down to nothing, and the manifold consists of two straight line segments. It is impossible to define the diffusion $\mathfrak{R}$ directly as the metric is singular at the ends of each line and l is discontinuous. The only possibility would be to consider a

sequence of smooth curves converging to the rectangle, and then considering the limit of the diffusions on the smooth curves. But even a circle will project down to two straight line segments and constructing the diffusions will be awkward because at the two singularities the drift is infinite.

However the diffusion Ω does have an easy interpretation. It is an Ornstein Uhlenbeck process on each line segment as l=constant here. When it hits the end, if it is on the lower (greater likelihood) segment, it will jump up to the upper segment if and only if it has enough energy (velocity) to do so. It will then start on the upper segment with a reduced velocity. If it has not enough energy to jump up then it will just reflect. If it hits the end of the upper segment it will always fall down and start with an increased velocity on the lower segment. This diffusion is simpler to construct as it hits the singularity only a finite number of times in a fixed time interval. It is also a strong solution which means that we have pathwise uniqueness for a given driving Brownian Motion. This follows because the Ornstein Uhlenbeck equation has a strong solution and there are just a finite number of paths to glue together.

■

Lyapunov Exponents

For the metrics h and $\tilde{h}$ where there is a natural embedding of M in $\mathbb{R}^K$ it is possible to start a diffusion at every point of M and then use the the same $\mathbb{R}^K$ Brownian Motion to generate each path.

This is called gradient Brownian Flow - see Carverhill[3]. For each t, $\Re(t)$ will be a smooth diffeomorphism of M (it is a given path evolving in t which is not differentiable).

In typical cases adjacent points will move together exponentially fast, and the rate is given by the Lyapunov Exponents. The easiest case to analyse is the exponential family under metric $\tilde{h}$, which makes the manifold Euclidean.

Exponential Family - Flat Metric

Suppose θ^α is the path starting at $\alpha \in \mathbb{R}^m$. Since the metric is independent of position in $\mathbb{R}^m$ the flow is given by:

$$d\theta_i^\alpha = \tilde{h}^{-1}_{ij}\left[\tau_j^{(r)} - \overline{\tau_j}\right]dB_r + \frac{1}{2}\tilde{h}^{-1}_{ij}\frac{\partial l}{\partial \theta_j}\,dt$$

Following Rogers[19] p141 we have

$$\theta_i^\alpha(t) = \tilde{h}^{-1}_{ik}\left[\tau_k^{(r)} - \overline{\tau_k}\right]B_r(t) + \frac{1}{2}\tilde{h}^{-1}_{ik}\int_0^t \frac{\partial l}{\partial \theta_k}(s)\,ds + \alpha_i$$

$$\frac{\partial \theta_i}{\partial \alpha_j} = \frac{1}{2}\tilde{h}^{-1}_{ik}\int_0^t \frac{\partial \theta_q}{\partial \alpha_j}\frac{\partial^2 l}{\partial \theta_q \partial \theta_k}(s)\,ds + \delta_{ij}$$

This is solved by $\frac{\partial \theta^\alpha}{\partial \alpha}(t) = \exp\left[\frac{1}{2}\tilde{h}^{-1}_{.j}\int_0^t \frac{\partial^2 l}{\partial \theta . \partial \theta_j}(s)\,ds\right]$ where exp acts on matrices.

Now recall that $l = \tau_k\theta_k - \psi(\theta)$ where $e^{\psi(\theta)} = \int e^{\tau_k\theta_k}\,dx$. So $-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j} = \frac{\partial^2\psi}{\partial\theta_i\partial\theta_j}$ and $\frac{\partial\psi}{\partial\theta_i} = e^{-\psi(\theta)}\int \tau_i e^{\tau_k\theta_k}\,dx = \mathbb{E}\left[\tau_i\right]$ and $\frac{\partial^2\psi}{\partial\theta_i\partial\theta_j} = \mathbb{E}\left[\tau_i\tau_j\right] - \mathbb{E}\left[\tau_i\right]\mathbb{E}\left[\tau_j\right]$ which is positive definite, as it is the covariance matrix of the natural statistics. Thus $\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}$ is negative definite for each θ and so $A_{sj} = \frac{1}{t}\int_0^t \frac{\partial^2 l}{\partial\theta_s\partial\theta_j}(s)\,ds$, the time-average of negative definite matrices, is negative definite. Since $\tilde{h}$ is positive definite, the matrix $D_{rs} = \tilde{h}^{-1}_{rj}A_{sj}$, which exp acts on above, is negative definite. This proves that

the Lyapunov Exponents are all negative and further that all points are moving together all the time. In general the concept refers to the long-time average behaviour of points starting sufficiently closely.

The actual exponents are given by the eigenvalues of $\underset{t\to\infty}{\text{Lim}}\ \frac{1}{t}\left[\frac{1}{2}\,\tilde{h}^{-1}_{ik}\int_0^t \frac{\partial^2 l}{\partial\theta_k\partial\theta_j}(s)\,ds\right]$. By the Ergodic theorem this equals $\mathbb{E}\left[\frac{1}{2}\,\tilde{h}^{-1}_{ik}\,\frac{\partial^2 l}{\partial\theta_k\partial\theta_j}\right]$ where the expectation is with respect to the stationary measure on M.

Now we consider the flow for the process Ω. Following the same method as before for $\begin{bmatrix}\theta\\V\end{bmatrix}$ starting at $\begin{bmatrix}\alpha\\\beta\end{bmatrix}$, we have:

$$\begin{bmatrix}\theta(t)\\V(t)\end{bmatrix} = \int_0^t \begin{bmatrix} V(s) \\ -2cV(s) + c\nabla l(\theta(s))\end{bmatrix}ds + \begin{bmatrix}\alpha\\\beta\end{bmatrix} + 2c\begin{bmatrix} 0 \\ \tilde{h}^{-1}_{\cdot j}\left[\tau_j(r) - \overline{\tau_j}\right]B_r\end{bmatrix}$$

$\frac{\partial(\theta,V)}{\partial(\alpha,\beta)} = \int_0^t \begin{bmatrix}0 & I\\ c\partial^2 l/\partial\theta^2 & -2cI\end{bmatrix}\frac{\partial(\theta,V)}{\partial(\alpha,\beta)}\,ds + I$ which is solved by:

$\frac{\partial(\theta,V)}{\partial(\alpha,\beta)} = \exp\int_0^t\begin{bmatrix}0 & I\\ -c\Sigma^{-1} & -2cI\end{bmatrix}ds$ where $\Sigma^{-1} = -\partial^2 l/\partial\theta^2$ is positive definite. Consider a fixed θ; by a rotation we can assume that $\Sigma = \text{diag}(\gamma_1^2,\ldots,\gamma_m^2)$. Then $\begin{vmatrix}-\lambda I & I\\ -c\Sigma^{-1} & (-2c-\lambda)I\end{vmatrix} = \prod\left[\lambda(2c+\lambda) + c\gamma_i^{-2}\right]$ so the eigenvalues are given by $\lambda(2c+\lambda) + c\gamma_i^{-2} = 0$. If $\gamma_i^2 > 1/c$ then the roots are complex with real part $-c$. If $\gamma_i^2 \leqslant 1/c$ the roots are given by $-c \pm \sqrt{c^2 - c\gamma_i^{-2}} < 0$.

A special case occurs when we have a multivariate normal distribution with unknown means but known covariance which is Σ (constant). A linear transformation ensures that this is diagonal so that γ_i^2 is the known variance in the i-direction.

Since $\frac{\partial(\theta,V)}{\partial(\alpha,\beta)} = e^{-At}$ for some constant matrix A we have that $\begin{bmatrix}\theta^{\alpha}(t) - \theta^{0}(t)\\ V^{\beta}(t) - V^{0}(t)\end{bmatrix}$ follows a deterministic path so that once we have one path, any other path converges towards it exponentially fast in a deterministic and smooth way.

Consider just the phase plane which gives the i-coordinate and i-velocity. If $\gamma_i^2 > 1/c$ then the plane is contracted by the flow at a uniform rate c, and also rotated at a constant rate. If $\gamma_i^2 \leqslant 1/c$ then the plane is shrunk uniformly at two rates in orthogonal directions.

Simulations of this behaviour are shown: four paths of different colours are started at the corners of the large square. FIGURE 2.1 is the flow $\mathfrak{R}$ with m=2; the paths are periodically joined together to show how the flow moves points towards each other in a smooth way. FIGURE 2.2 is the flow Ω with m=1; position is plotted horizontally and velocity is plotted vertically. The square is shrunk and rotated by the flow.

To picture the general exponential model, this type of behaviour occurs locally, but as the path moves around the manifold, so the type of shrinking changes. It is the drift which pushes adjacent paths together while the brownian term does nothing as it affects all paths equally.

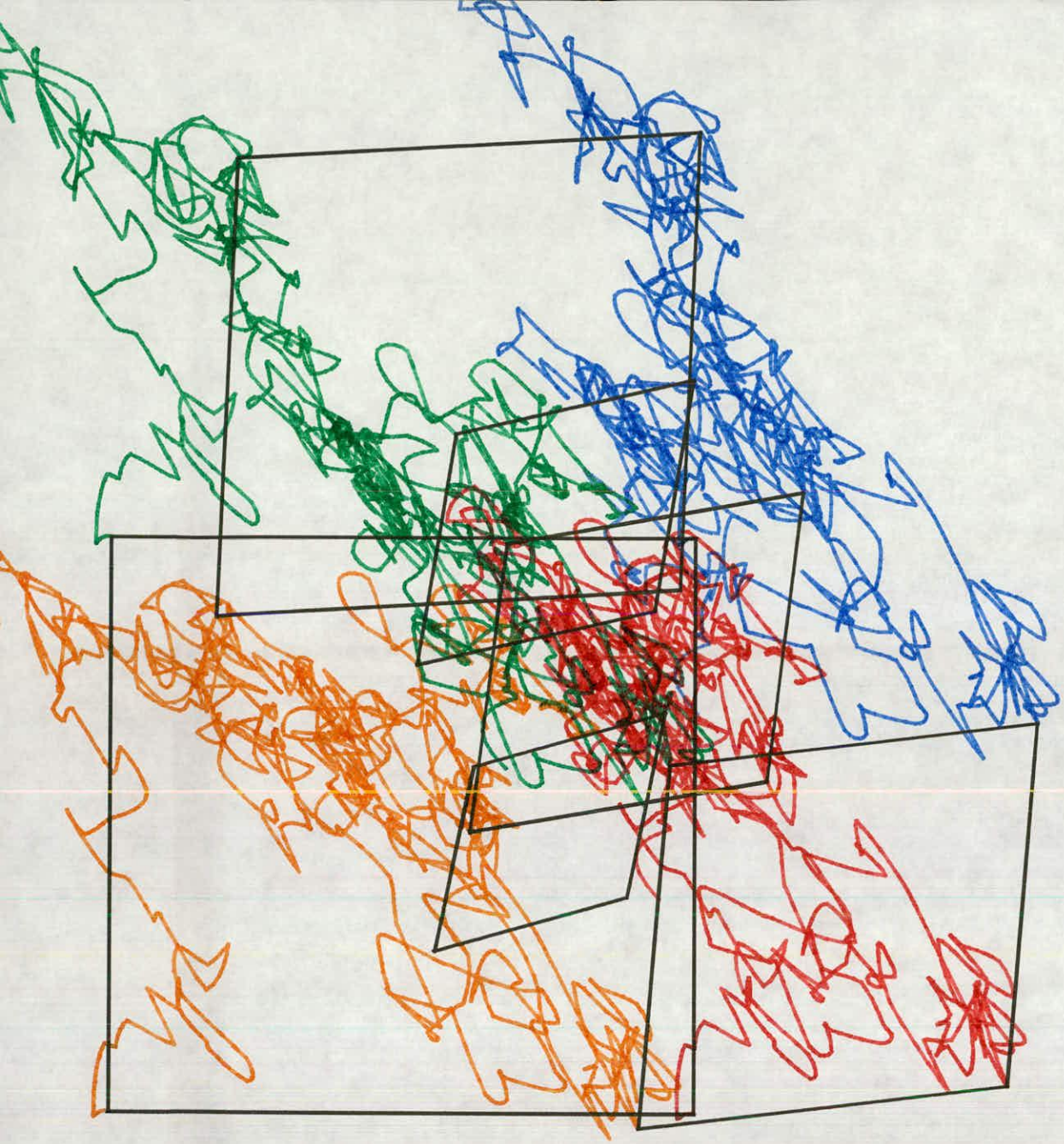
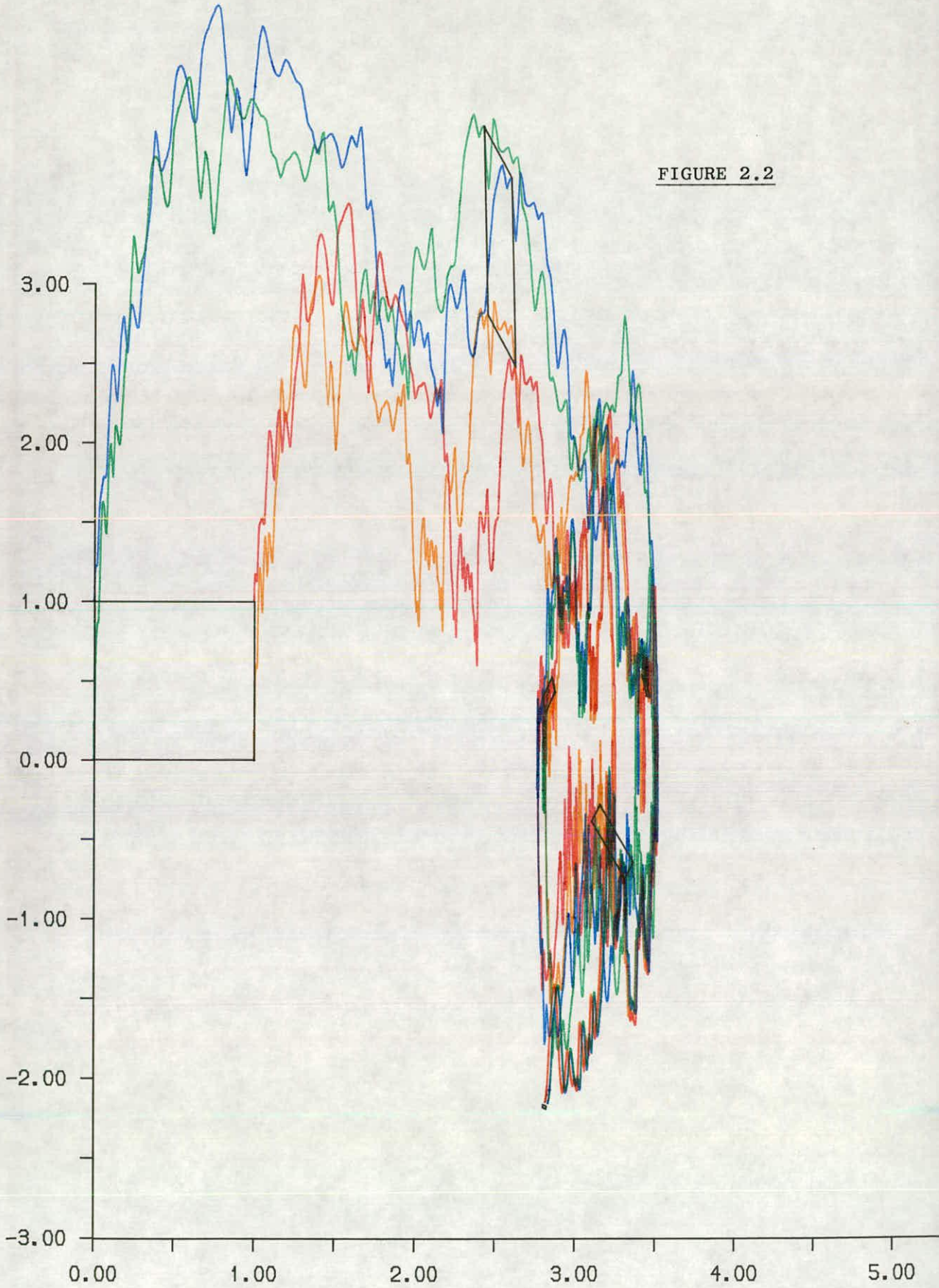


FIGURE 2.1

FIGURE 2.2



General Case - Standard Embedding

Now consider the general case under the empirical metric h, which gives the more natural embedding. Firstly note that considering Riemannian Manifolds derived from a Statistical Model is in fact no limitation. A theorem of Nash and Gromov states that any m-dimensional Riemannian manifold can be embedded in $\mathbb{R}^{m(m+1)/2+3m+5}$. Given a manifold embedded in some $\mathbb{R}^q$, it will be possible to find a model so that the coordinate functions are the log-likelihoods for certain observations.

One-Dimensional Case

When the manifold is curved it is a surprising property of gradient brownian flow that adjacent points will be pushed together without any drift - see Carverhill[3]. In one dimension the intuitive result is that the log of the distance between two points starting close together is a brownian motion with a negative drift of half the square of the curvature of the manifold at that point (to first order). As the example of an embedded circle shows, the flow must always be a diffeomorphism so that points cannot be simultaneously pushed together everywhere on the manifold.

Now consider the drift which is one half the gradient of the total log-likelihood. If the 1-dimensional manifold is convex with respect to increasing likelihood then the drift will push adjacent points apart. If it is concave then it will push them together - see FIGURE 2.3. It can be deduced from the calculations in Rogers[19] p141 that (to first order) the log of

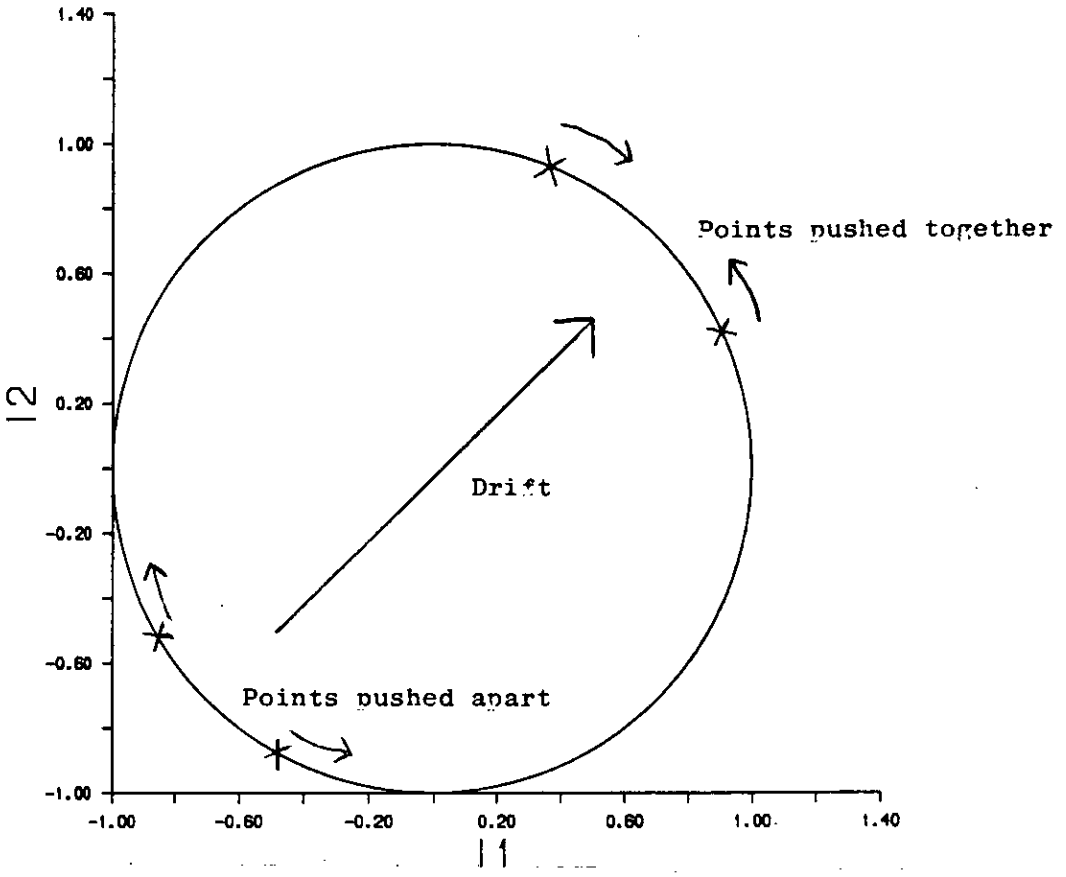


FIGURE 2.3

the distance between two close points is increasing at a rate of the dot product of the curvature and the total drift which is $\frac{1}{2}(1,1,\ldots,1)^T$.

Let the manifold be $\{l_1(\theta),\ldots,l_\kappa(\theta)\} \subset \mathbb{R}^\kappa$, where θ is arc length so that $g = (\partial l_i/\partial\theta)^2 = 1$. Then gradient brownian flow with drift is given by

$$d\theta = \frac{\partial l_r}{\partial\theta}\, dB_r + \frac{1}{2}\sum\frac{\partial l_r}{\partial\theta}\, dt \quad \text{(see above)}$$

If θ^α is the path starting at α then

$$\theta^\alpha(t) = \int\frac{\partial l_r}{\partial\theta}\, dB_r(s) + \frac{1}{2}\sum\frac{\partial l_r}{\partial\theta}\, ds$$

$$\frac{\partial\theta^\alpha}{\partial\alpha} = \int\frac{\partial^2 l_r}{\partial\theta^2}\,\frac{\partial\theta}{\partial\alpha}\, dB_r(s) + \frac{1}{2}\sum\frac{\partial^2 l_r}{\partial\theta^2}\,\frac{\partial\theta}{\partial\alpha}\, ds$$

This sde has the explicit solution

$$\frac{\partial\theta^{\alpha}}{\partial\alpha} = \exp\left[\int \frac{\partial^2 l_r}{\partial\theta^2}\, dB_r(s) + \frac{1}{2}\sum \frac{\partial^2 l_r}{\partial\theta^2} - \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2 ds\right]$$

since by Ito's Formula this implies that

$$d\frac{\partial\theta^{\alpha}}{\partial\alpha} = \frac{\partial\theta^{\alpha}}{\partial\alpha}\left[\frac{\partial^2 l_r}{\partial\theta^2}\, dB_r(s) + \frac{1}{2}\sum \frac{\partial^2 l_r}{\partial\theta^2} - \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2 ds\right] + \frac{1}{2}\frac{\partial\theta^{\alpha}}{\partial\alpha}\sum\left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2 ds$$

The Lyapunov Exponent is given by

$$Lim\ \frac{1}{t}\int_{s=0}^{t} \frac{\partial^2 l_r}{\partial\theta^2}\, dB_r(s) + \frac{1}{2}\sum \frac{\partial^2 l_r}{\partial\theta^2} - \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2 ds$$

Now if the flow is ergodic, the first term will be 0. This is because $\int_{s=0}^{t} \frac{\partial^2 l_r}{\partial\theta^2}\, dB_r(s)$ is a time-changed brownian motion, and the time it has run for is $T = \int_{s=0}^{t} \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2 ds$. By the properties of brownian motion, its value is $O(\sqrt{T})$. By ergodicity, $T/t \longrightarrow \mathbb{E}\left[\left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2\right]$ (assumed finite) where $\mathbb{E}$ is expectation with respect to the stationary measure of the diffusion.

By ergodicity the Lyapunov Exponent will be $\mathbb{E}\left[\frac{1}{2}\sum \frac{\partial^2 l_r}{\partial\theta^2} - \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2\right]$. So:

$$LE = \int_M \left[\frac{1}{2}\sum \frac{\partial^2 l_r}{\partial\theta^2} - \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2\right] e^{\Sigma l_r} d\theta \Big/ \int_M e^{\Sigma l_r} d\theta$$

Integration by parts, assuming that either the manifold is compact or that $\Sigma l_r \longrightarrow -\infty$ as $\theta \longrightarrow \pm\infty$ gives:

$$LE = -\frac{1}{2}\int_M \left[\left[\sum \frac{\partial l_r}{\partial\theta}\right]^2 + \left[\frac{\partial^2 l_r}{\partial\theta^2}\right]^2\right] e^{\Sigma l_r} d\theta \Big/ \int_M e^{\Sigma l_r} d\theta < 0$$

A diffusion on the interest parameter space

At the moment the metric and log-likelihood are defined on M. To make inferences about the interest parameter we need functions defined on N. Given the diffusion $\Re(t)$ then $\varphi(\Re)$ is a process on N, but it is not Markov in general as the increments will depend on the position of $\Re$ in M.

A metric on N

For $\theta \in M$ and $\lambda \in N$ with $\varphi(\theta) = \lambda$ we have an inner product on TN_λ called the orthogonalized information, $g^{\perp}$ - see Amari p251. This is defined on two vectors in TN_λ by lifting them to the unique vectors in TM_θ which are orthogonal to the fibre, and then using the original metric. In coordinates the orthogonalized information matrix has dual given by the top left corner of $g^{-1}_{ij}(\theta)$. The orthogonalized information is the only sensible choice which does not depend on how the fibres are parameterized. To get a metric dual on N, we average this quantity over the fibre above λ with respect to some probability distribution on the fibre. This gives us the *harmonic mean metric*, g^{T} on N. Some justification for considering this is the following: suppose we have $\zeta : N \longrightarrow \mathbb{R}$ and we want an expression for its gradient at λ. The only natural route is:

$\zeta \circ \varphi : M \longrightarrow \mathbb{R}$

$\nabla \zeta \circ \varphi \ (\theta) \in TM_\theta$

$\varphi' \nabla \zeta \circ \varphi \ (\theta) \in TN_\lambda$

Define $\tilde{\nabla} \zeta \ (\lambda) = \int_{\varphi^{-1}(\lambda)} \varphi' \nabla \zeta \circ \varphi \ (\theta) \ d\beta$ where $d\beta$ represents a probability distribution on the fibre.

Lemma: $\tilde{\nabla}$ is the gradient operator on N with respect to harmonic mean metric.

Proof: $\varphi' \nabla \zeta \circ \varphi \ (\theta) = (I \ 0) g^{-1}_{ij} \begin{bmatrix} \partial \zeta / \partial \lambda \\ 0 \end{bmatrix}$

■

Suppose the interest parameter space is 1-dimensional. Then if the interest parameter is the first coordinate the quantity we

wish to average over the fibre is g_{11}^{-1}. (This refers to the 1,1 coordinate of g^{-1}.) If λ is the interest parameter this is equal to $< \nabla\lambda, \nabla\lambda >$. Expressing it in this way can be helpful if λ is not one of the coordinates used to specify M.

Consider again g as an inner product on the tangent space at a point of M. If λ_i are interest parameters then the subspace spanned by the $\nabla\lambda_i$ will be perpendicular to the fibre. The restriction of g to this subspace gives $g^{\perp}$, the orthogonalized information. The orthogonal complement of this subspace will be parallel to the fibre and g restricted to this subspace will be $g^{\parallel}$. Of course $g=g^{\perp}+g^{\parallel}$. Now if we replace g with $\bar{g} = g^{\perp}+g^{\parallel}$, the effect is to average the distance between adjacent fibres, thus making them 'parallel'. Gradient under the metric $\bar{g}$ is denoted $\bar{\nabla}$.

Now that we have a metric on the interest parameter space N we can put a Brownian Motion on it, but to obtain a drift we need a vector field on N which corresponds to the vector field ∇l on M. So for $\lambda \in N$ we need a vector in TN_λ. Now for $\theta \in M$, $\varphi'\bar{\nabla}l(\theta) \in TN_\lambda$ where $\varphi(\theta) = \lambda$ (see FIGURE 2.4). As θ ranges over $\varphi^{-1}(\lambda)$, $\varphi'\bar{\nabla}l(\theta)$ is always a vector in TN_λ so we can average these vectors with respect to the probability distribution on $\varphi^{-1}(\lambda)$ used above.

Note at this stage that if λ is a 1-dimensional interest parameter then the quantity we wish to average is $\bar{g}(\bar{\nabla}l, \bar{\nabla}\lambda) = \bar{g}_{ij}\,\bar{g}^{ik}\,\frac{\partial l}{\partial \theta_k}\,\bar{g}^{jm}\,\delta_{m1} = \bar{g}^{1k}\,\frac{\partial l}{\partial \theta_k} = \bar{\nabla}l\Big]_1 = \varphi'\bar{\nabla}l(\theta)$, where superscript implies g^{-1}.

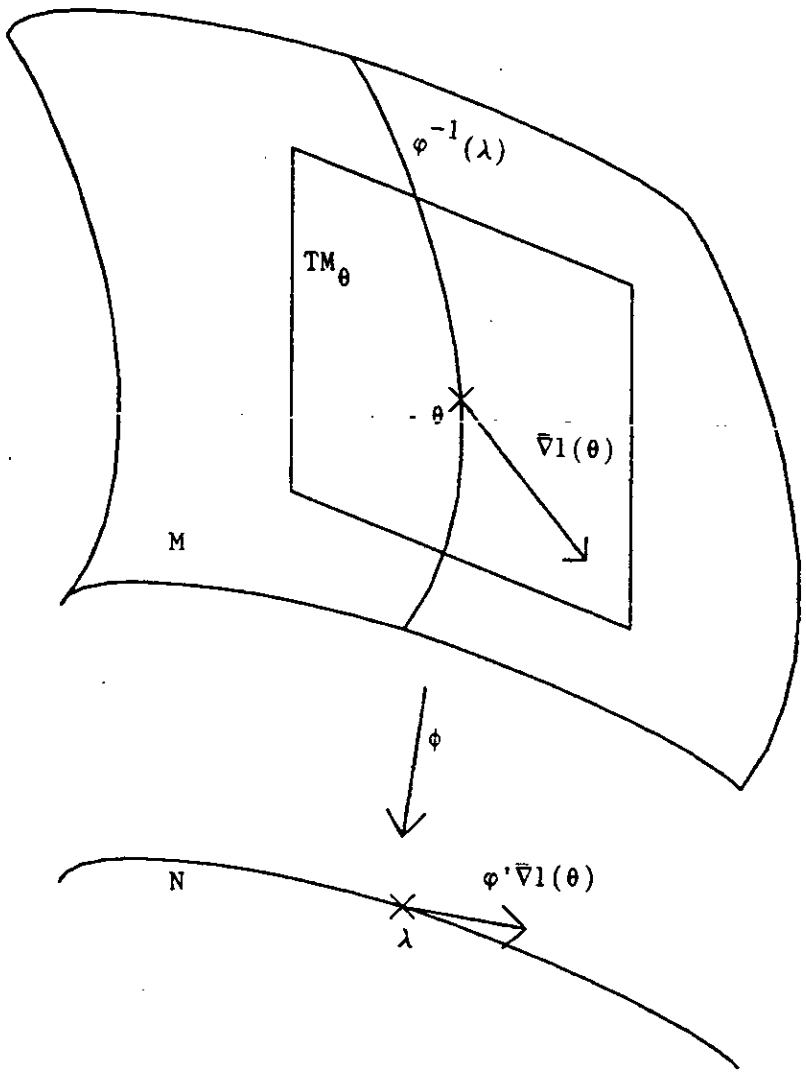

$\varphi^{-1}(\lambda)$
TM_θ
θ
$\bar{\nabla} l(\theta)$
M
ϕ
N
$\varphi' \bar{\nabla} l(\theta)$
λ

FIGURE 2.4

The choices that are still left are: which metric to use, and which probability distribution to put on the fibres. The best choices will vary from problem to problem and this will be illustrated in subsequent chapters. The vector field and metric downstairs can now be used as a basis for inference of the interest parameter. The most convenient way to do this is to find the stationary measure of the diffusion which is brownian motion on N with drift one half the vector field. Basing inference for the interest parameter on this metric and this vector field shall be called projection method. The next chapter gives some motivation for this procedure by considering the case of incidental parameters where standard procedures can fail.

It should be pointed out that since the vector field and the metric were derived separately, there is no reason why the vector field should be a gradient with respect to the metric. A one-dimensional vector field is always a gradient of some function, but in more than one dimension the vector field need not be a gradient with respect to the particular metric derived. In three dimensions this will be the case if the vector field has non-zero curl, and therefore tends to push the particle around in circles. However we can still derive a likelihood measure for the interest parameter by taking the stationary measure for the diffusion on the interest parameter space. This will have to be done by simulation in general as a formula exists only when the drift is a gradient.

Chapter 3. Inference in the presence of incidental parameters

The concept of incidental and structural parameters was introduced by Neymann and Scott [17] and can be stated as follows. There is a probability law depending on two parameters, λ and β. For $i=1,\ldots,\kappa$, X_i is a random variable with law specified by λ and β_i. Thus λ is the structural parameter fixed throughout and the β_i are incidental parameters which are different for each observation. Any of λ,β,X may be vector-valued. The object is inference about the structural parameter λ. It can be helpful to assume that the β_i are selected from an unknown distribution ν, independently from the X_i. This makes the problem semi-parametric - the unknowns are a parameter λ and a distribution ν.

Kalbfleisch and Sprott [13] give two cases where the incidental parameters can be systematically eliminated, without losing any information on the interest parameter, in terms of factorizing the density:

I) Marginal case: $f(X;\lambda,\beta)\partial(u,a)/\partial(X) = f(a;\lambda)f(u|a;\lambda,\beta)$ where $X = (a,u)$ and the second factor contains no available information concerning λ in the absence of knowledge of β.

II) Conditional Case: $f(X;\lambda,\beta) = f(X|T;\lambda)f(T;\lambda,\beta)$ where the second factor contains no available information concerning λ in the absence of knowledge of β.

In case I inference for the interest parameter λ should be based entirely on the observations a_i and the observations u_i ignored. In case II inference for the interest parameter λ should be based on the fact that we have independent observations X_i with density $f(.|T_i;\lambda)$ and the actual distribution of the T_i should be ignored.

These two cases are genuinely different; they could be combined for a unified theory by considering;

$$f(X;\lambda,\beta,\gamma) = f(u|a,T;\lambda,\beta)f(a|T;\lambda)f(T;\lambda,\beta)$$

but this unnecessarily complicates the notation.

The best way to enforce the rather subjective rider on available information is to use the concept of sufficiency - see Sprott[20]. For sufficiency in the marginal case there has to be an irreducible pivotal relation between β and $u|a$. The density of u conditional on a is given by $f(u|a;\lambda,\beta)$. We need to find a function $H(\beta,u,a)$, the pivot, whose distribution conditional on a does not depend on β, though it may depend on λ. The irreducibility requirement means that H must be one-one in u for fixed β. This will be illustrated in *example 2*. Intuitively the observations u_1 and u_2 will give rise to the same information about λ since for each β_1 there is a β_2 with (u_1,β_1) and (u_2,β_2) giving rise to the same conditional likelihood for λ. Further details on pivots are given below.

In the conditional case, we need an irreducible pivotal relation between T and β.

Background on Pivotal Distributions

Pivotal or Fiducial distributions were invented by R.A. Fisher in an attempt to quantify belief in different regions of the parameter space on the basis of the results of a single experiment. The difference from the confidence region approach is that the latter considers repeating the experiment many times and the long term proportion of success. In many situations the experiment is only performed once, so what might have occured is not strictly relevant. Unfortunately, pivotal distributions are not easy to interpret satisfactorily. As an example of the derivation of a pivotal distribution, consider the case of one observation from a location model.

Location Parameter Case: If μ is a location parameter then the density for random variable X is $f_X(x) = \rho(x-\mu)$. The quantity $H = X-\mu$ has a distribution which does not depend on the parameter μ: $f_H(y) = \rho(y)$. Such a quantity H is called a pivot. The pivot H is a random variable because X is, while μ is a fixed (but unknown) constant. Once the observation x has been made, we can fix X at this value and then artificially give μ a distribution so that H remains distributed as before. Here we write μ_0 for the fixed unknown value and μ for the random variable with the pivotal distribution.

$\mathbb{P}[\mu < c] = \mathbb{P}[x-\mu > x-c]$ since x is the fixed observation.

$= \mathbb{P}[X-\mu_0 > x-c]$ from pivoting

$= 1 - \mathbb{P}[H < x-c] = 1 - \int_{-\infty}^{x-c} \rho(y)dy.$

So the density for μ is $f_\mu(c) = \rho(x-c)$.

Scale Parameter Case: $f_X(x) = \frac{1}{\sigma}\rho(x/\sigma)$

$H = X/\sigma$ so $f_H(y) = \rho(y)$

$\mathbb{P}[\sigma < c] = \mathbb{P}[x/\sigma > x/c] = \mathbb{P}[H > x/c]$ from pivoting

$= \int_{x/c}^{\infty} \rho(y)dy.$

So differentiating with respect to c: $f_\sigma(c) = \rho(x/c).x/c^2$ is the pivotal density.

■

Lemma: With one observation from the location case and the scale case, the likelihood measure under the Fisher Metric equals the pivotal measure.

Proof:

Location Case: The pivotal measure is simply the likelihood with respect to Lebesgue Measure. The Fisher Information for a location parameter is constant since $\mathbb{E}\left[\frac{\partial}{\partial\mu}\log\rho(X-\mu)\right]^2$ is expectation of a function of $X-\mu$ only. So under the Fisher Metric a location parameter is a Euclidean parameter.

Scale Case: The Fisher Information, $\mathbb{E}\left[\frac{\partial}{\partial\sigma}\log\left[\frac{1}{\sigma}\rho(X/\sigma)\right]\right]^2 = \frac{1}{\sigma^2}\mathbb{E}\left[1 + \frac{(X/\sigma)\rho'(X/\sigma)}{\rho(X/\sigma)}\right]^2 = \frac{\text{constant}}{\sigma^2}.$

Therefore the likelihood measure is proportional to $\frac{1}{\sigma}\rho(x/\sigma)\sqrt{(1/\sigma^2)}d\sigma$ which is the pivotal measure. [Recall that the likelihood function is only defined up to a multiplicative constant].

■

When the state space and the parameter space are equal to $\mathbb{R}$, the pivotal distribution is well-defined and unique, as the distribution function itself can be taken as the pivot. However

it is equal to the likelihood measure only in special cases. Where there are more than one observation pivotals do not necessarily exist and are not necessarily unique if they do. A full discussion is given in Wilkinson[23]; conditions are based on factorizing the density.

A convenient interpretation of a pivotal distribution is the following: suppose M the parameter space and $\mathfrak{X}$ the sample space are homeomorphic; let θ_0 and x_0 be fixed points in the parameter and sample spaces; suppose further that $X:\Omega \times M \longrightarrow \mathfrak{X}$ is a homeomorphism for each $\omega \in \Omega$, which depends on ω only through $X(\omega,\theta_0)$. The observation is $x = X(\omega,\theta_0)$. By the conditions above we can find a $\tilde{\theta} \in M$ (depending on x,x_0,θ_0) which satisfies $X(\omega,\tilde{\theta}) = x_0$ for the same ω. Then $\tilde{\theta}$ will have the pivotal distribution given observation x_0. So for the existence of a pivot we need at least some sort of duality between the sample space and a the parameter space.

Example 3.1: Suppose $X_i \sim N(\mu_i,\sigma^2)$ with the X_i independent, which means we are in the marginal case with $\beta=\mu$, $\lambda=\sigma^2$, $u=X$ and $a=\emptyset$. The pivotal distribution for μ_i is $N(X_i,\sigma^2)$. Since the density simply defines a pivotal relation between X and μ, there is no information left for inference about σ^2 according to sufficiency considerations.

Looking at this as a semi-parametric problem, we have iid observations from the density $G = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \nu(\mu)d\mu$ where σ^2 and ν are unknown.

$$\mathbb{E}[X^2] = \int\int \frac{x^2}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\ \nu(\mu)d\mu\ dx$$

$= \mathbb{E}[\mu^2] + \sigma^2$ (assuming ν is smooth).

So $\mathbb{V}[X] = \sigma^2 + \mathbb{V}[\mu] \geqslant \sigma^2$.

This means that we will become increasingly confident about an upper bound for σ^2 as more observations are taken; if σ_1^2 and σ_2^2 are less than $\sigma_0^2 + \mathbb{V}[\mu]$, where σ_0^2 is the true value, the data will not be able to distinguish between σ_1^2 and σ_2^2 however many observations are taken.

This model can be interpreted in terms of a diffusion process. Let Z_t be a brownian motion with initial density ν. Then at time σ^2, the density will be G. Suppose we actually knew the form of G, which is the most that could be recovered however many observations are taken. Then there is no way of telling exactly how long the diffusion has been running for: for $\tau < \sigma^2$, it could have started at time τ with density $\int \frac{1}{\sqrt{2\pi\tau}}\, e^{-(x-\mu)^2/2\tau}\ \nu(\mu)d\mu$ and then run for a further time $\sigma^2 - \tau$.

■

Example 3.2: Suppose $X_i \sim N(\mu,\sigma_i^2)$ independently. The pivotal distribution for $1/\sigma^2$ is $\Gamma\left[\frac{(X_1-\mu)^2}{2}\ ,\ \frac{1}{2}\right]$. This is reducible since $\text{sign}(X-\mu)$ has a distribution independent of σ^2.

Considering this as a semi-parameteric problem, the density is $\int \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\ \nu(\sigma^2)d\sigma^2$. The expectation of X will exist if:

$$\int_{x=0}^{\infty}\int \frac{x}{\sqrt{2\pi\sigma^2}}\, e^{-x^2/2\sigma^2}\ \nu(\sigma^2)d\sigma^2 = \int \sqrt{\frac{\sigma^2}{2\pi}}\ \nu(\sigma^2)d\sigma^2 < \infty.$$

In this case the sample mean of the X_i will be a consistent estimator for μ. This example does not fall into i) or ii) above.

■

Example 3.3: (See Neymann and Scott [17].)

X_i, $Y_i \sim N(\mu_i,\sigma^2)$, $i=1,\ldots,\kappa$, all 2κ observations independent. Here we have that $a=(X-Y)/2$, $u=(X+Y)/2$, $f(x,y;\mu,\sigma^2) \propto f(a;\sigma^2)f(u;\mu,\sigma^2)$. The second term defines an irreducible pivotal relation between u and μ viz $(u-\mu) \sim N(0,\sigma^2/2)$. Thus consideration of u will give no assistance in finding σ^2, so inference for σ^2 should be based entirely on a.

The maximum likelihood estimators are given by $\hat{\mu}_i = (X_i + Y_i)/2$ and $\hat{\sigma}^2 = \Sigma(X_i - Y_i)^2/4\kappa$ which is an inconsistent estimator for σ^2. In fact as $\kappa \longrightarrow \infty$, $\hat{\sigma}^2 \longrightarrow \sigma^2/2$. The reason that things are going wrong is that we are restricting attention to just one value of μ_i, when with only X_i and Y_i relating to μ_i there is no justification for this. It may be that it is $\mu_1 = (x_1+y_1)/2$ which makes the observations x_1 and y_1 most likely, but it is not reasonable to assume that all the random variables $(X_i+Y_i)/2$ are actually equal to their means. It would seem a much better policy not to narrow each μ_i down to a single value, but to find a likelihood distribution for μ_i in the light of the observation.

■

If we use the Fisher Metric to give a volume element to the parameter space then the likelihood with respect to this volume element is the likelihood measure. We could then take the

marginal for the interest parameter, and we seem to have considered a range of values for the nuisance parameter. Unfortunately this method fails in general as *example 3.3* shows:

Example 3.3 cont. The Fisher metric is $\mathrm{Diag}(\kappa/\sigma^4, 2/\sigma^2, \ldots, 2/\sigma^2)$. The volume element is proportional to the square root of the determinant of this matrix ie to $(\sigma^2)^{-1-\kappa/2} d\sigma^2 d\mu_1 \ldots d\mu_\kappa$. So the likelihood measure is proportional to $(\sigma^2)^{-1-\kappa/2-\kappa} e^{-\Sigma(x_i-\mu_i)^2/2\sigma^2-(y_i-\mu_i)^2/2\sigma^2} d\sigma^2 d\mu_1 \ldots d\mu_\kappa =$ $(\sigma^2)^{-1-\kappa-\kappa/2} e^{-\Sigma(\mu_i-[x_i+y_i]/2)^2/\sigma^2-(x_i-y_i)^2/4\sigma^2} d\sigma^2 d\mu_1 \ldots d\mu_\kappa$.

Integrating out the μ_i leaves:

$(\sigma^2)^{-1-\kappa} e^{-\Sigma(x_i-y_i)^2/4\sigma^2} d\sigma^2 \propto \theta^{\kappa-1} e^{-\theta\Sigma(x_i-y_i)^2/4} d\theta$

where $\theta=1/\sigma^2$, ie

$$\theta = 1/\sigma^2 \sim \Gamma[\Sigma(x_i-y_i)^2/4,\ \kappa].$$

This is an inconsistent estimating measure as $\Sigma(x_i-y_i)^2/4\kappa \longrightarrow \sigma^2/2$.

■

Return now to the general model with incidental parameters in case I or II above. If we have a genuine prior for the nuisance parameters then we could use Bayesian techniques to find a posterior for the interest parameter. In the iid case Bayesian techniques are fairly robust to misspecifying the prior, provided there is a large amount of data. In the incidental parameter case the errors associated with misspecifying the prior will remain however large the sample is. As an example of this consider the following:

Example 3.4:

The structural parameter is p with prior U(0,1).

The incidental parameter $q_i \in [0,1]$ has prior density ν.

$X_i=1$ with probability pq_i and 0 otherwise.

The posterior is therefore proportional to:

$$\prod pq_i^{X_i}(1-pq_i)^{1-X_i}\nu(q_i)dq_i \quad dp$$

If $X_i=1$ exactly α out of κ times then integrating out the q_i gives $(p\mathbb{E}[q])^{\alpha}(1-p\mathbb{E}[q])^{\kappa-\alpha}dp$, where $\mathbb{E}$ is expectation under ν.

As $\kappa \longrightarrow \infty$, $\alpha/\kappa \longrightarrow \lambda$, a constant and the posterior will tend to a delta function at $\lambda/\mathbb{E}[q]$. Misspecification of ν will lead to misspecification of $\mathbb{E}[q]$. This will lead to an incorrect posterior tending to a delta function at the wrong value.

■

The following section shows how projection method can be applied. The probability distribution to put on the fibres of constant interest parameter will be the pivotal distribution, ρ. Consider first the Marginal case so that $f(x;\lambda,\beta) \propto f(u|a;\lambda,\beta)f(a;\lambda)$.

$$\bar{\nabla}\log f(x;\lambda,\beta) = \bar{\nabla}\log f(u|a;\lambda,\beta) + \bar{\nabla}\log f(a;\lambda).$$

The projection of the second component, $\varphi'\bar{\nabla}\log f(a;\lambda) \in TN_{\lambda}$ does not depend on position in the fibre because $\bar{\nabla}$ was constructed to be independent of position on the fibre and $f(a;\lambda)$ is independent of β. So the average of this component over the fibre will be $\varphi'\bar{\nabla}\log f(a;\lambda)$. We will show that integrating the first component over the fibre gives 0. Let u_0 be the observation and β_0 the true parameter (fixed) while β has the pivotal distribution defined by u_0, and u is random with distribution defined by β_0. Then we show that $\log f(u;\beta_0)$ and $\log f(u_0;\beta)$ are equal in distribution. We

know that $\mathbb{E}[\frac{\partial}{\partial\lambda}\log f(u;\beta_0)] = 0$ so by using linearity $\mathbb{E}[\varphi'\bar{\nabla}\log f(u_0;\beta)] = 0$. But this expectation is simply integrating $\varphi'\bar{\nabla}\log f(u_0;\beta)$ over the fibre with respect to pivotal measure.

Let $H=H(\beta_0,u,a)$ be the pivot so that the distribution of H conditional on a is independent of β_0. The coordinate for the fibre, β must be chosen so that H does not depend explicitly on λ. We must choose a metric so that the vectors $\partial/\partial\beta$ and $\partial/\partial\lambda$ are orthogonal. Let H have density $f_H(\alpha)$. So the pivotal density for β is given by $\rho d\beta = f_H(\alpha)\frac{\partial H}{\partial\beta}\,d\beta$ and $f(u|a;\beta) = f_H(\alpha)\frac{\partial H}{\partial u}$.

The vectors $\partial/\partial\lambda$ and $\partial/\partial\beta$ orthogonal means that on the fibre the matrix which gives the inner product can be written $\begin{bmatrix}\bar{g}_{11} & 0\\ 0 & g_{22}\end{bmatrix}$ where $\bar{g}_{11}$ is the part of the inner product matrix which covers the span of the $\partial/\partial\lambda$. Since this is already an average over the fibre, it is a function of λ and maybe X only. Then $\varphi'\bar{\nabla}\log f(u|a;\beta) = \bar{g}_{11}^{-1}\left[\frac{\partial}{\partial\lambda}\log f_H(\alpha) + \frac{\partial}{\partial\lambda}\log\frac{\partial H}{\partial u}\right]$. But H is independent of λ (though its distribution may depend on λ). So the second term is 0. We know that $\mathbb{E}\left[\frac{\partial}{\partial\lambda}\log f_H(\alpha)\right] = 0$ so integrating $\varphi'\bar{\nabla}\log f(u|a;\beta)$ over the fibre with respect to pivotal measure is bound to give 0.

Example 3.3 - continued: In this case, $g = \bar{g}$ since the metric depends only on the interest parameter σ^2. The pivotal measure for μ_i is $N\left[\frac{x_i+y_i}{2}, \frac{\sigma^2}{2}\right]$. The score statistic, $\nabla\log f(x,y) = g^{-1}\begin{bmatrix}-\kappa/\sigma^2+1/2\sigma^4\Sigma[(x_i-\mu_i)^2+(y_i-\mu_i)^2]\\ 1/\sigma^2[x_i-\mu_i+y_i-\mu_i]\\ \vdots\end{bmatrix}$.

The projected score statistic, $\varphi'\nabla \log f(x,y)$ is simply the first component which is $\sigma^4\{-1/\sigma^2+1/2\kappa\sigma^4\Sigma[(x_i-\mu_i)^2 + (y_i-\mu_i)^2]\}$ $= -\sigma^2+1/\kappa\sum\left[\left[\mu_i-\frac{x_i+y_i}{2}\right]^2+\left[\frac{x_i-y_i}{2}\right]^2\right]$. Averaging this with respect to pivotal measure gives $-\sigma^2+1/\kappa\sum\left[\frac{\sigma^2}{2}+\left[\frac{x_i-y_i}{2}\right]^2\right] = -\frac{\sigma^2}{2}+\frac{1}{\kappa}\sum\left[\frac{x_i-y_i}{2}\right]^2$ (*).

Recall that the metric downstairs is $\bar{g}(\partial/\partial\sigma^2,\partial/\partial\sigma^2) = \kappa/\sigma^4$. The function with (*) as a gradient is $-\frac{\kappa}{2}\log(\sigma^2) - \frac{1}{\sigma^2}\sum\left[\frac{x_i-y_i}{2}\right]^2$, and so the likelihood measure for σ^2 is $(\sigma^2)^{-\kappa/2}\ e^{-\Sigma(x_i-y_i)^2/4\sigma^2}d\sigma^2/\sigma^2$ which can be written $1/\sigma^2 \sim \Gamma(\Sigma(x_i-y_i)^2/4\ ,\ \kappa/2)$. This measure is consistent and is what would have been obtained by basing inference on the X_i-Y_i. However it was not necessary to find the factorization explicitly, only to find a pivotal distribution on the nuisance fibres. This also gives motivation for how to tackle more complex distributions.

Generalization

This result gives motivation for an approach to the general problem when no pivot exists. The problem is to find a suitable likelihood measure for the fibres $\varphi^{-1}(\lambda)$. For the case of incidental parameters sampled from an unknown measure ν, we consider the sample as being iid with unknown parameters λ, the interest parameter and the measure ν. Kiefer and Wolfowitz[15] showed that under certain regularity conditions, the values of λ and ν which maximize the likelihood are consistent estimators.

In general finding them is a very hard problem: the likelihood is $\prod\int f(x_i;\lambda,\beta_i)\nu(\beta_i)d\beta_i$, and we have to choose the density ν which maximizes this. In practice the only feasable approach, even allowing for numerical techniques, is to restrict the class of possible ν's under consideration. Many different cases are explored in Maritz and Lwin[16]. Once we have an estimate $\hat{\nu}$ for ν, then $\prod f(x_i;\lambda,\beta_i)\hat{\nu}(\beta_i)d\beta_i$ gives the likelihood measure on the fibre above λ. In this case we have used the data to find the best estimate for the prior measure on the fibres; in the pivotal case, we did not attempt to estimate the prior, but instead used the model to give a measure on the fibres which would automatically ignore extraneous information.

Lemma: Suppose we know ν, the prior measure on the fibres. Then, for a particular choice of metric, the averaged vector field on the interest parameter space produced by projection method is the same as the vector field derived when the problem is considered to be iid sampling from the density $\int f(X;\beta,\lambda)\nu(\beta)d\beta$.

Given ν, we can choose a coordinate system β on the nuisance fibre is such that the density $\nu(\beta)$ is independent of λ. The metric chosen must be such that $\partial/\partial\beta$ and $\partial/\partial\lambda$ are orthogonal.

Proof: Using existing notation, $\varphi'\bar{\nabla}\log f(x;\lambda,\beta) = \bar{g}_{11}^{-1}\left[\frac{\partial \log f(x;\beta,\lambda)}{\partial\lambda}\right]$. So the vector in TN_λ obtained by averaging is

$$\sum_{i=1}^{K} \frac{\int \bar{g}_{11}^{-1}\left[\frac{\partial \log f(x_i;\beta_i,\lambda)}{\partial\lambda}\right] f(x_i;\beta_i,\lambda)\nu(\beta_i)d\beta_i}{\int f(x_i;\beta_i,\lambda)\nu(\beta_i)d\beta_i}$$

$$= \bar{g}_{11}^{-1}\sum \frac{\partial}{\partial\lambda}\log\int f(x_i;\beta,\lambda)\nu(\beta)d\beta, \text{ as } \nu \text{ is independent of } \lambda.$$

■

Now if we can find a sequence of measures $\hat{\nu}_K$ which tend weakly to the true prior ν, then we will have $\mathbb{E}\left[\frac{\partial}{\partial\lambda}\log\int f(X;\beta,\lambda)\hat{\nu}_K(\beta)d\beta\right] \longrightarrow 0$ and so our averaged vector in TN_λ will tend to 0 at the true λ.

Under the conditions of [15] this procedure always gives a consistent estimating measure for θ. However there are some surprising cases when the crucial identifiability condition fails.

The actual density is $\int f(x;\beta,\lambda)\nu(\beta)d\beta$. If we could find $\lambda_1 \neq \lambda_2$ and ν_1, ν_2 with $\int f(x;\beta,\lambda_1)\nu_1(\beta)d\beta = \int f(x;\beta,\lambda_2)\nu_2(\beta)d\beta$, then there is no way we will be able to decide between λ_1 and λ_2 however many observations are taken. The simplest case is the following example:

Example 3.5: Integrated circuits are produced independently. Each one functions correctly with a fixed probability p. The circuits are given a quick test by a machine which lets through some junk and the probability it does so depends on the nature of the fault. So for circuit i, there is a probability $q_i(1-p)$ that it is rejected and a probability $(1-p)(1-q_i)$ that it is let through and is junk. A second test then sorts out the remaining junk. The process is modelled by assuming that the q_i are independently selected from an unknown

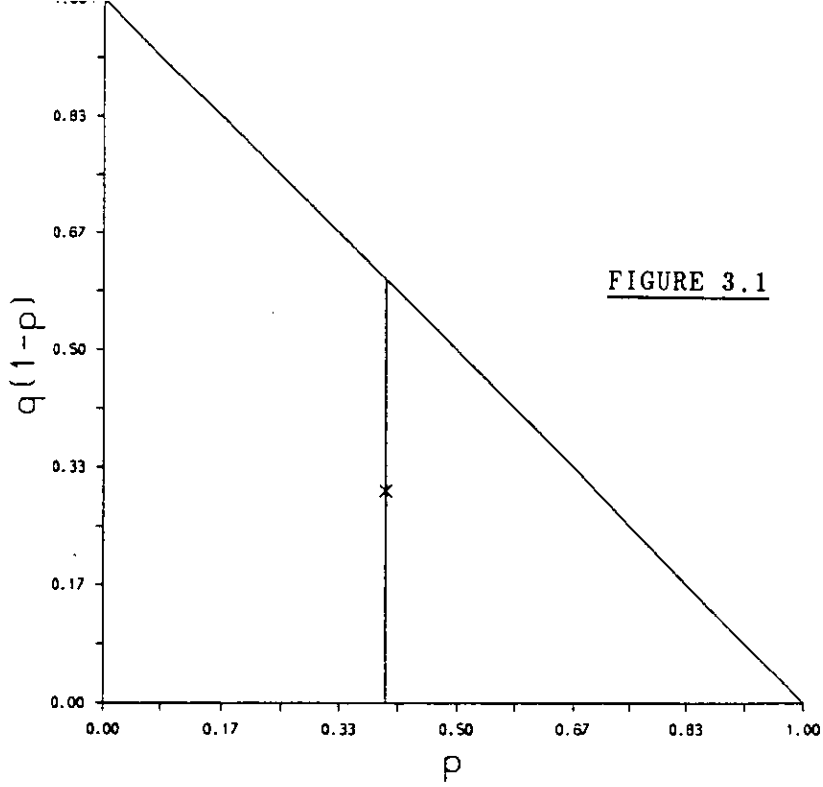


FIGURE 3.1

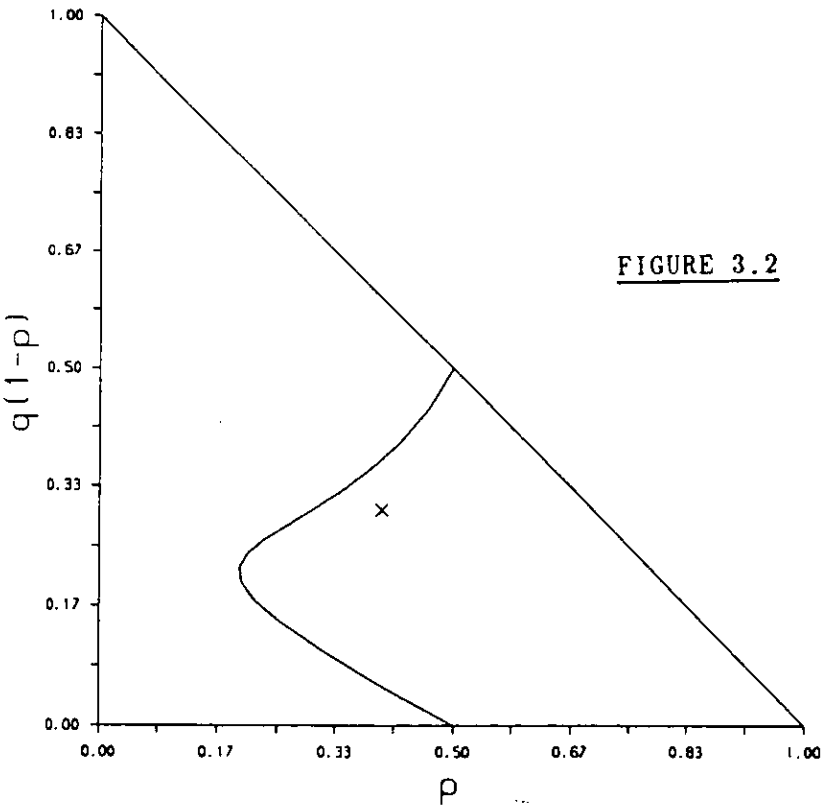


FIGURE 3.2

distribution ν. It is easy to see that it is equivalent to assume that q_1 is fixed at $q=\mathbb{E}_\nu(q_1)$.Both p and q can be consistently estimated but not ν. Considering the parameter space as a 2-simplex notice that each value of p corresponds to a straight line (see FIGURE 3.1). The measure ν is supported on one of these lines. The model is equivalent if ν is replaced by point mass at its centre of mass. Since the line is straight the centre of mass of any measure supported on that line must lie on the same line. But now suppose that the line of constant interest parameter were curved (see FIGURE 3.2). Then as ν varies the centre of mass describes the convex hull of the line. All that can ever be inferred from the observations is this centre of mass. However this will be contained in the convex hulls of different interest parameters. Hence the interest parameter is non-identifiable.

■

The same identification failure occurs in the following more surprising example.

Example 3.6: Let X_1,Y_1 be iid with density $(1+2\alpha x+3\beta x^2)/(1+\alpha+\beta)dx$, $\alpha > 0$, $\beta > 0$, $0 \leqslant x \leqslant 1$. Thus the joint density is $\dfrac{(1+2\alpha(x+y)+4\alpha^2xy+3\beta(x^2+y^2)+6\alpha\beta(xy^2+yx^2)+9\beta^2x^2y^2)dxdy}{(1+\alpha+\beta)^2}$

The parameter space is $(\mathbb{R}^+)^2$ and the interest parameter will be $\theta = \beta e^{-\alpha}$. Thus the unknowns are the $\theta \in \mathbb{R}^+$ and a probability measure on the line $\beta = \theta e^{\alpha}$. Now if we were to consider a convex combination of normal densities:

$\int \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-(x-\mu)^2/2\sigma^2} \nu(\mu,\sigma^2)d\mu d\sigma^2$ then the space of such densities would be infinite dimensional, as is the space of kernel densities ν. However the convex combination:

$$\int \frac{(1+2\alpha(x+y)+4\alpha^2xy+3\beta(x^2+y^2)+6\alpha\beta(xy^2+yx^2)+9\beta^2x^2y^2)}{(1+\alpha+\beta)^2} \nu(\alpha,\beta)d\alpha d\beta$$

depends only on the quantities $1-\eta = \int \frac{\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, $\eta_1 = \int \frac{2\alpha\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, $\eta_2 = \int \frac{\alpha^2\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, $\eta_3 = \int \frac{2\beta\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, $\eta_4 = \int \frac{2\alpha\beta\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, $\eta_5 = \int \frac{\beta^2\nu(\alpha,\beta)d\alpha d\beta}{(1+\alpha+\beta)^2}$, where the scaling is such that $\eta=\eta_1+\eta_2+\eta_3+\eta_4+\eta_5$. All the η's must lie between 0 and 1 so the set of convex combinations of the original densities gives a density of the following form which lies in the 5-simplex: $(1-\eta+\eta_1(x+y)+4\eta_2xy+\frac{3}{2}\eta_3(x^2+y^2)+3\eta_4(xy^2+yx^2)+9\eta_5x^2y^2)dxdy$. If the parameter is first selected from a certain measure on this 5-simplex and then x and y generated with this parameter, then by linearity the observed random process would be the same if the parameter were fixed at the centre of gravity of the measure. The actual parameter space is the set of (α,β) which is a 2-dimensional submanifold of the 5-simplex, given by $\left\{\left[\frac{2\alpha}{(1+\alpha+\beta)^2}, \frac{\alpha^2}{(1+\alpha+\beta)^2}, \frac{2\beta}{(1+\alpha+\beta)^2}, \frac{2\alpha\beta}{(1+\alpha+\beta)^2}, \frac{\beta^2}{(1+\alpha+\beta)^2}\right] : 0 < \alpha < 1,\ 0 < \beta < 1\right\}$. The lines $\beta=\theta e^{\alpha}$ foliate the parameter space so for each value of the interest parameter θ there is a line in the parameter space. The tangent vector is given by:

$$\frac{1}{(1+\alpha+\theta e^{\alpha})^3}\begin{bmatrix} 2(1+\alpha+\theta e^{\alpha}) - 4\alpha(1+\theta e^{\alpha}) \\ 2\alpha(1+\alpha+\theta e^{\alpha}) - 2\alpha^2(1+\theta e^{\alpha}) \\ 2\theta e^{\alpha}(1+\alpha+\theta e^{\alpha}) - 4\theta e^{\alpha}(1+\theta e^{\alpha}) \\ 2\theta e^{\alpha}(1+\alpha)(1+\alpha+\theta e^{\alpha}) - 4\alpha\theta e^{\alpha}(1+\theta e^{\alpha}) \\ 2\theta^2 e^{2\alpha}(1+\alpha+\theta e^{\alpha}) - 2\theta^2 e^{2\alpha}(1+\theta e^{\alpha}) \end{bmatrix}$$

We now show that each line has a convex hull which contains an open set in the 5-simplex. The only way this can fail is if the line is restricted to a 4-plane within the 5-simplex. To see that this does not happen simply look at the tangents at 5 different points and check that they span the whole of 5-space. This is done most easily by putting them side by side to form a 5×5 matrix and numerically checking that this is non-singular. A more satisfactory way to check that the line is always twisting sufficiently for its convex hull to be space-filling is to consider the first four derivatives of the tangent vector (wrt α) and see that these form a non-singular matrix; this approach requires more calculation, but can be tackled by computer. The package *Mathematica* gives the determinant as :

$$\frac{8\theta^4 e^{4\alpha}}{(1+\alpha+\theta e^{\alpha})^5}\left[(2\alpha-5)^2+9\right]$$

By continuity, given any fibre there will be a neighbouring fibre such that the corresponding hulls intersect. The most that can be inferred from the observations, however many are made, is the centre of mass of ν over the true fibre. From the above consideration there will be θ_1,ν_1 and θ_2,ν_2 both giving the same centre of mass so the parameters are non-identifiable.

■

The surprising fact about *example 3.6* is that the sample space has the same dimension (2) as the parameter space, and the observation appears to give useful information on both parameters. It is not a trivial or exceptional example as general foliations (instead of $\beta=\theta e^{\alpha}$) would give the same result. I do not know of any general conditions on the density for identifiability of the parameters beyond factorization of the form I) or II) being sufficient.

Chapter 4 Further Applications

The ideas described in the previous section can be easily extended to other situations. In the usual case of independent sampling with a fixed number of parameters, it is still a useful technique to find an estimating measure for the nuisance parameters and then produce a mean vector field and a harmonic mean metric on the interest parameter space. If the sample size is small it is a bad idea to estimate the nuisance parameters by their mle's as restricting consideration to just one value could be misleading. Proving any sort of optimality will be hard because pivotal distributions do not exist in general.

As before, λ is the interest parameter, β the nuisance parameter and $X_1,\ldots,X_\kappa$ the observations. For a pivotal distribution we need the decomposition of $X_1,\ldots,X_\kappa$ into statistics y and z so that

$\prod_1 f(x_i;\lambda,\beta)\frac{\partial(y,z)}{\partial(x)} = f(y,z;\lambda,\beta) = f(y|z;\beta,\lambda)f(z;\lambda)$ where the first term on the right defines an irreducible pivotal relation between $y|z$ and β, and further that this pivot is independent of λ - see Wilkinson[23]. The examples considered here do not come into this exceptional category. In the absence of any prior information a sensible measure to put on the fibres is the likelihood measure normalized to be a probability measure on the fibre. This depends on the metric.

We will end up with a vector field and a metric on the interest parameter space N. From these, a likelihood measure can be constructed which represents our belief in certain areas of the parameter space. This will depend on the metric originally

chosen so it is important to be sure that this metric will yield sensible results. It can be useful to put the diffusion straight onto the interest parameter space and then use the fact that its stationary measure is this likelihood measure. This is because simulation of the diffusion only requires knowledge of the vector field and the metric. This is especially helpful when the parameter space has dimension greater than 1 so that the likelihood measure cannot be graphed easily, and when the vector field is not a gradient so that the likelihood measure cannot be calculated explicitly. Suppose λ_1 and λ_2 are both interest parameters. Then our belief that λ_1 were greater than λ_2 is given by $meas\{N:\lambda_1 > \lambda_2\}/meas\{N\}$, where *meas* is the likelihood measure on the interest parameter space N. By the ergodic theorem the proportion of time that the λ_1-coordinate of the diffusing particle is greater than the λ_2-coordinate will tend to this value, and this can be a simple way of quantifying belief.

Timekeepers Example

Suppose there are j=1,...,J timekeepers each recording the times T_{ij} of i=1,...,I athletes to the nearest 1/100 sec. The true times μ_i are unknown. We assume that $T_{ij} = [X_{ij}]$, the integer part of X_{ij} for normal independent random variables X_{ij}. We also assume that the timekeepers are unbiased but that each has his own variance σ_j^2. These variances are of interest. A possible difficulty is that in this situation the overall likelihood is unbounded if we set $\mu_i = x_{i1}$ as $\sigma_1^2 \rightarrow 0$. This means that as we tend towards the degenerate case when the first timekeeper is

perfect, at the subset of the parameter space where $\mu_i = x_{i1}$, the first timekeeper's observations, the likelihood due to the first timekeeper tends to infinity; however the likelihoods due to the other timekeepers stay bounded away from 0 so the overall likelihood tends to infinity. Looking at the likelihood alone is clearly unsatisfactory.

In these coordinates the Fisher Information Matrix is diagonal with $g(\partial/\partial\mu_i, \partial/\partial\mu_i) = 1/\sigma_1^2 + \ldots + 1/\sigma_J^2$, $g(\partial/\partial\sigma_j^2, \partial/\partial\sigma_j^2,) = I/2\sigma_j^4$. This only depends on the interest parameters so equals $\bar{g}$. We also have $\partial l/\partial\sigma_j^2 = -I/2\sigma_j^2 + \Sigma_i(x_{ij}-\mu_i)^2/2\sigma_j^4$.

The measure derived from the metric on a fibre of constant σ^2 is constant (ie independent of μ) which corresponds to each μ_i having a uniform prior over the range of interest. This would be a reasonable assumption anyway. So the posterior density for μ_i is proportional to $\exp\left[-\sum \frac{(x_{ij}-\mu_i)^2}{2\sigma_j^2}\right] d\mu_i$ ie $\mu_i \sim N\left[\frac{\Sigma x_{ij}/\sigma_j^2}{\Sigma\ 1/\sigma_j^2}\ ,\ \frac{1}{\Sigma\ 1/\sigma_j^2}\right]$, where the sums are over j.

The projected vector field $\varphi'\bar{\nabla}l$ has jth component $-\sigma_j^2 + \Sigma_i\ (x_{ij}-\mu_i)^2/I$, and taking the expectation of this under the above distribution for μ_i gives:

$$-\sigma_j^2 + \frac{1}{I}\sum x_{ij}^2 - 2x_{ij}\frac{\Sigma x_{ik}/\sigma_k^2}{\Sigma\ 1/\sigma_k^2} + \left[\frac{\Sigma x_{ik}/\sigma_k^2}{\Sigma\ 1/\sigma_k^2}\right]^2 + \frac{1}{\Sigma\ 1/\sigma_k^2}.$$

Finally note that $X_{ij}|T_{ij} \sim U(T_{ij}, T_{ij}+1)$ approximately so that the final score statistic downstairs is:

$$D_j = -\sigma_j^2 + \frac{1}{I}\sum t_{ij}^2 + t_{ij} + \frac{1}{3} - \frac{2/\sigma_j^2}{12\Sigma 1/\sigma_k^2} - 2(t_{ij}+\frac{1}{2})\frac{\Sigma(t_{ik}+1/2)/\sigma_k^2}{\Sigma 1/\sigma_k^2} + \frac{\Sigma\ 1/\sigma_k^4}{12(\Sigma 1/\sigma_k^2)^2} + \left[\frac{\Sigma(t_{ik}+1/2)/\sigma_k^2}{\Sigma 1/\sigma_k^2}\right]^2 + \frac{1}{\Sigma\ 1/\sigma_k^2}$$

The best way to make use of this is to simulate a diffusion which is brownian motion (with respect to the metric) plus drift D/2 on the interest parameter space, $(\mathbb{R}^+)^J$. The coordinate representation is:

$$d\sigma_j^2 = \sigma_j^2\sqrt{2/I}\ dB_j + \frac{1}{2}\left[-\frac{2\sigma_j{}^2}{I} + \frac{4\sigma_j{}^2}{I} + D_j\right]dt$$

where B is an $\mathbb{R}^J$ Brownian motion - see Chapter 2.

The results for 4 timekeepers timing 42 races are given below. We estimate the σ's under the likelihood measure by sampling the diffusion at every time interval up to time 100 and taking the mean. The belief that Timekeeper i had a smaller variance than timekeeper j was also estimated from the diffusion for each pair i,j. Results are given for 5 sets of test data with s1=20, s2=25, s3=30, s4=35, and real data taken from a BAL competition at Luton in 1989.

Data	s1	s2	s3	s4	1<2	1<3	1<4	2<3	2<4	3<4
1	24	28	35	58	.58	.80	.97	.63	.96	.86
2	19	29	37	40	.77	.91	.90	.64	.76	.52
3	22	18	27	32	.28	.69	.82	.73	.94	.62
4	22	37	36	35	.86	.81	.82	.44	.44	.47
5	24	26	32	35	.68	.74	.78	.55	.71	.54
Real	26	31	26	17	.68	.52	.20	.26	.11	.20

In this particular case a more usual way of getting rid of the μ_i's would be to consider differences between timekeepers and thus obtain estimates ξ_{ij} for sums $\sigma_i^2 + \sigma_j^2$. While this is bound to give consistent estimates, there is no guarantee that for a

small data set the estimate for σ_1^2 will be positive. For J=3 the estimate will be $(\xi_{12}+\xi_{13}-\xi_{23})/2$. Using projection method we are estimating from a diffusion on the interest parameter space, and the estimates will therefore always be permissible.

Example - The Behrens Fisher Problem.

This example concerns two sets of data from different normal distributions.

$X_i \overset{iid}{\sim} N(\mu_1,\sigma_1^2)$ $i=1,\ldots,\kappa$

$Y_j \overset{iid}{\sim} N(\mu_2,\sigma_2^2)$ $j=1,\ldots,\chi$, independently from the X_i.

The mean difference $\nu = \mu_1 - \mu_2$ is of interest.

The log-likelihood

$$l = -\frac{\kappa}{2}\log[\sigma_1^2] - \Sigma(x_i-\mu_1)^2/2\sigma_1^2 - \frac{\chi}{2}\log[\sigma_2^2] - \Sigma(y_j-\mu_2)^2/2\sigma_2^2.$$

The Fisher Metric $g = diag\{\kappa/\sigma_1^2,\kappa/2\sigma_1^4,\chi/\sigma_2^2,\chi/2\sigma_2^4\}$ for the parameters taken in order $\mu_1,\sigma_1^2,\mu_2,\sigma_2^2$.

$$< \nabla\nu,\nabla\nu > = (1,0,-1,0)diag\{\sigma_1^2/\kappa,2\sigma_1^4/\kappa,\sigma_2^2/\chi,2\sigma_2^4/\chi\}(1,0,-1,0)^T$$

$$= \sigma_1^2/\kappa + \sigma_2^2/\chi.$$

From chapter 2 the measure on the parameter space is given by

$1/\sigma_1^2 \sim \Gamma(\kappa s_1/2,\kappa/2)$, $\mu_1|\sigma_1^2 \sim N(\bar{x},\sigma_1^2/\kappa)$,

$1/\sigma_2^2 \sim \Gamma(\kappa s_2/2,\chi/2)$, $\mu_2|\sigma_2^2 \sim N(\bar{y},\sigma_2^2/\chi)$

where $s_1=\Sigma(x_i-\bar{x})^2/\kappa$. and $s_2=\Sigma(y_j-\bar{y})^2/\chi$. The measure on the nuisance fibre is therefore this measure conditional on $\mu_1 - \mu_2 = \nu$.

Let $\psi = \mu_1 + \mu_2$. The measure on the nuisance fibre is given by ρ, the joint density of σ_1^2 and σ_2^2 (depending on ν) and the distribution of $\psi|\sigma_1^2,\sigma_2^2,\nu$ which is normal with mean $\bar{\psi} = \bar{x} + \bar{y} + (\sigma_1^2\chi - \sigma_2^2\kappa)(\nu + \bar{y} - \bar{x})/(\sigma_1^2\chi + \sigma_2^2\kappa)$ and variance $4\sigma_1^2\sigma_2^2/(\sigma_1^2\chi + \sigma_2^2\kappa)$.

$l = -\frac{\kappa}{2}\log[\sigma_1^2] - \Sigma(x_i - \frac{\psi+\nu}{2})^2/2\sigma_1^2 - \frac{\chi}{2}\log[\sigma_2^2] - \Sigma(y_j - \frac{\psi-\nu}{2})^2/2\sigma_2^2.$

$\frac{\partial l}{\partial \nu} = \Sigma\,(x_i - \frac{\psi+\nu}{2})/2\sigma_1^2 - \Sigma(y_j - \frac{\psi-\nu}{2})/2\sigma_2^2.$

This averages to $\Sigma\,(x_i - \frac{\bar{\psi}+\nu}{2})/2\sigma_1^2 - \Sigma(y_j - \frac{\bar{\psi}-\nu}{2})/2\sigma_2^2.$

$= \kappa\bar{x}/2\sigma_1^2 - \chi\bar{y}/2\sigma_2^2 - \nu(\kappa/4\sigma_1^2 + \chi/4\sigma_2^2) + \bar{\psi}(\chi/4\sigma_2^2 - \kappa/4\sigma_1^2)$

$= (\bar{x}-\bar{y}-\nu)\chi\kappa/(\chi\sigma_1^2+\kappa\sigma_2^2).$

So the quantities we need to average over σ_1^2 and σ_2^2 are $\sigma_1^2/\kappa + \sigma_2^2/\chi$ to get the metric dual and $(\bar{x}-\bar{y}-\nu)\chi\kappa/(\chi\sigma_1^2+\kappa\sigma_2^2)$ to get the score statistic.

Because of the complicated form of ρ we cannot proceed further analytically. Numerically, we need to integrate $\left[\sigma_1^2/\kappa + \sigma_2^2/\chi\right]\rho(\sigma_1^2,\sigma_2^2)$ over the space of $\{\sigma_1^2,\sigma_2^2\}$. This is not easy because ρ does not have a simple form and is only known up to a normalizing constant. An alternative is to use simulation. If we could simulate a sequence of pairs $\left[\sigma_1^2(i),\sigma_2^2(i)\right]$ from ρ then we could estimate $\int\left[\sigma_1^2/\kappa + \sigma_2^2/\chi\right]\rho(\sigma_1^2,\sigma_2^2)d\sigma_1^2d\sigma_2^2$ by simply averaging $\sigma_1^2(i)/\kappa + \sigma_2^2(i)/\chi$ over i. Any direct method of simulation such as rejection will be awkward to implement because of the form of ρ. Instead suppose we decide to calculate the score statistic and metric at $\nu = -4,-3.8,-3.6,\ldots,4$. Then we simulate $\sigma_1^2,\sigma_2^2,\mu_1,\mu_2$ from their simple unconditional distribution. Then if $\nu=\mu_1-\mu_2=$ 2.36 (say) take that σ_1^2,σ_2^2 as a simulation from $\rho|\nu=2.4$.

This is an efficient scheme as no simulations are wasted. Once we have a score statistic on the space of ν we can numerically integrate to get a derived log-likelihood. The exponential of this gives the derived likelihood. The following figures show the likelihood measure for ν which is obtained by multiplying the derived likelihood by the volume element and normalizing so that it integrates to 1.

FIGURE 4.1: $\kappa=30;\ \chi=30;\ \mu_1=2;\ \mu_2=4;\ \sigma_1^2=2;\ \sigma_2^2=4$

FIGURE 4.2: $\kappa=10;\ \chi=50;\ \mu_1=2;\ \mu_2=2;\ \sigma_1^2=0.5;\ \sigma_2^2=3$

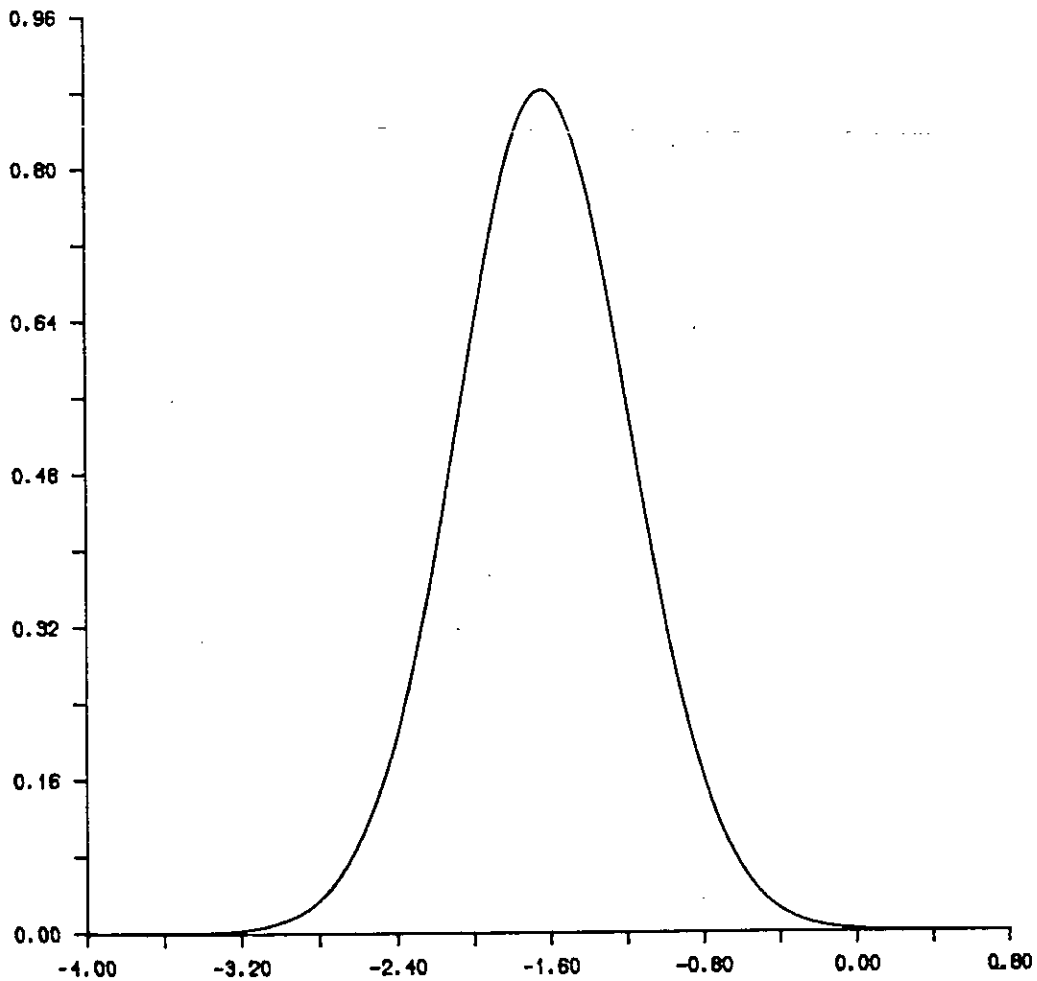


FIGURE 4.1

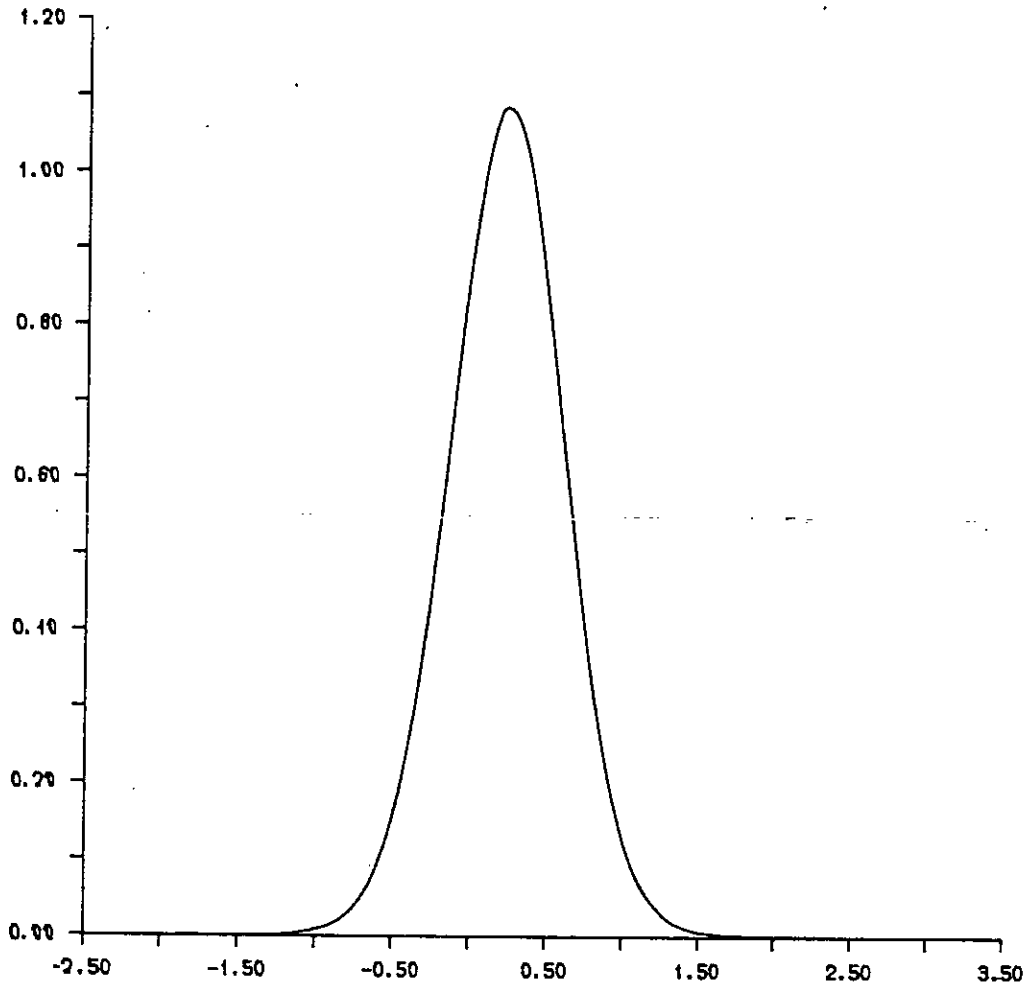


FIGURE 4.2

Mixture Model Example:

With the Behrens Fisher Problem we know which observations come from which distribution. An interesting variant is when all we know is that an observation comes from the first distribution with probability p and from the second with probability 1-p. Thus the observations are iid with density:

$$(2\pi)^{-1/2}\left[\frac{p}{\sigma_1}\exp\left[\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sigma_2}\exp\left[\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right]\right]$$

where $\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,p$ are the parameters. This type of distribution is typically awkward to work with and has the additional problem that at $\mu_1 = x_1$ the first observation, for any σ_2^2,p and μ_2 the likelihood is unbounded as $\sigma_1^2 \longrightarrow 0$. This is because it is always possible that the first observation came from the first distribution and all the rest came from the second distribution. Then the likelihood is maximized by setting $\mu_1 = x_1$ and letting $\sigma_1^2 \longrightarrow 0$. Relying on the mle is clearly useless so our method involves integrating over the parameter space and ensuring that the contribution from a neighbourhood of $\sigma_1^2 = 0$ is negligible. This requires using the right metric; since $\sigma_1^2 = 0$ is distant from the true value the performance of the Fisher and Empirical metrics will be fundamentally different.

To simplify the problem we reduce the number of parameters to two while maintaining the essence of the problem (see Cox[4] p291). Suppose that the random variable X is selected from $N(\mu,1)$ with probability 1/2 and from $N(\mu,\sigma^2)$ with probability 1/2. Thus X has a 2-parameter distribution with density:

$f(x) = \frac{1}{2\sqrt{2\pi}}\left(e^{-(x-\mu)^2/2} + \frac{1}{\sigma} e^{-(x-\mu)^2/2\sigma^2}\right)$. Take n iid observations; since this is a location model, without loss of generality these can be taken as $0, x_1, \ldots, x_{n-1}$. The likelihood function, $f(0)f(x_1)\ldots f(x_{n-1})$ is then unbounded at $\mu=0$ as $\sigma^2 \rightarrow 0$. The term in the expansion of the likelihood function which is going to cause problems is the one representing 0 being selected from $N(\mu,\sigma^2)$ and $x_1,\ldots,x_{n-1}$ being selected from $N(\mu,1)$ ie $e^{-\Sigma(x_i-\mu)^2/2}.\sigma^{-1}e^{-\mu^2/2\sigma^2}$. In fact it is sufficient to look at $\sigma^{-1}e^{-\mu^2/2\sigma^2}$ as the rest is constant to first order in a neighbourhood of $(0,0)$.

Note that $\frac{\partial l}{\partial \mu} = \frac{(x-\mu)\left[e^{-(x-\mu)^2/2} + \sigma^{-3}e^{-(x-\mu)^2/2\sigma^2}\right]}{e^{-(x-\mu)^2/2} + \sigma^{-1}e^{-(x-\mu)^2/2\sigma^2}}$,

$$\frac{\partial l}{\partial \sigma^2} = \frac{(-1/2\sigma^3+(x-\mu)^2/2\sigma^5)e^{-(x-\mu)^2/2\sigma^2}}{e^{-(x-\mu)^2/2} + \sigma^{-1}e^{-(x-\mu)^2/2\sigma^2}}.$$

If we had a true prior, the resulting posterior would be bound to be a true distribution and using this would be a sensible way to proceed. In the absence of a true prior we could choose to use the measure derived from the metric as an (improper) prior on the parameter space and then consider the 'posterior' measure. This resulting measure fails to integrate for both the Fisher and empirical metrics.

i) Fisher metric (Jeffreys Prior).

$$g_{11} = \mathbb{E}\left(\frac{\partial \log f(X)}{\partial \mu}\right)^2$$
$$= \int \frac{1}{8\pi}(x-\mu)^2\left[e^{-(x-\mu)^2/2} + \sigma^{-3}e^{-(x-\mu)^2/2\sigma^2}\right]^2 \frac{f(x)dx}{f^2(x)}$$

$$= \int \frac{1}{\sqrt{8\pi}} x^2 \frac{\left[e^{-x^2/2} + \sigma^{-3}e^{-x^2/2\sigma^2}\right]^2}{\left[e^{-x^2/2} + \sigma^{-1}e^{-x^2/2\sigma^2}\right]} dx$$

$$> \int \frac{1}{\sqrt{8\pi}} x^2 \frac{\left[e^{-x^2/2} + \sigma^{-3}e^{-x^2/2\sigma^2}\right]^2}{\left[1 + \sigma^{-1}\right]} dx \quad = O(1/\sigma^2)$$

since the critical term is:

$$\sigma^{-6}/\sigma^{-1}\int x^2 \, e^{-x^2/\sigma^2} \, dx.$$

$$g_{12} = \mathbb{E}\left(\frac{\partial \log f(X)}{\partial \mu} \cdot \frac{\partial \log f(X)}{\partial \sigma^2}\right) = 0.$$

$$g_{22} = E\left(\frac{\partial \log f(X)}{\partial \sigma^2}\right)^2 = \int \frac{1}{8\pi}\left[\frac{-1}{2\sigma^3} + \frac{(x-\mu)^2}{2\sigma^5}\right]^2 e^{-(x-\mu)^2/\sigma^2} \cdot \frac{f(x)dx}{f^2(x)}$$

$$= \int \frac{1}{\sqrt{8\pi}} \cdot \frac{(-1/2\sigma^3 + x^2/2\sigma^5)^2 e^{-x^2/\sigma^2}}{\left[e^{-x^2/2} + \sigma^{-1}e^{-x^2/2\sigma^2}\right]} dx$$

$$> \int \frac{1}{\sqrt{8\pi}} \cdot \frac{(1/4\sigma^6 - x^2/2\sigma^8 + x^4/4\sigma^{10})}{1 + 1/\sigma} e^{-x^2/\sigma^2} \, dx = O(1/\sigma^4).$$

So the volume element of the Fisher metric g is $\sqrt{|g|}d\mu d\sigma^2 > O(1/\sigma^3)d\mu \, d\sigma^2$. Consider integrating the likelihood function with respect to this measure over a neighbourhood of (0,0).

Recall that the essential part of the likelihood function is: $\sigma^{-1}e^{-\mu^2/2\sigma^2}$

Let $0 < \eta < \epsilon$.

$$\int_\eta^\epsilon d\mu \int_0^{\epsilon^2} d\sigma^2 \, \sigma^{-1}e^{-\mu^2/2\sigma^2} \left[\frac{1}{\sigma^3}\right] \qquad [w=1/\sigma^2]$$

$$= \int_\eta^\epsilon d\mu \int_{\epsilon^{-2}}^{\infty} \cdot \frac{dw}{w^2} \, e^{-w\mu^2/2} \, w^2$$

$$= \int_\eta^\epsilon d\mu \; 2/\mu^2 \; e^{-\mu^2/2\epsilon^2} \longrightarrow \infty \text{ as } \eta \longrightarrow 0.$$

ii) Empirical metric.

For simplicity, assume there are two observations: 0,x. Let l_1 and l_2 be the two log-likelihoods. Then $|h| = \left[\frac{\partial l_1}{\partial \mu} \cdot \frac{\partial l_2}{\partial \sigma^2} - \frac{\partial l_1}{\partial \sigma^2} \cdot \frac{\partial l_2}{\partial \mu}\right]^2$.

$$\sqrt{|h|} \qquad =$$

$$\frac{1}{8\pi f(x)f(0)}\left|-\mu\left[e^{-\mu^2/2} + \sigma^{-3}e^{-\mu^2/2\sigma^2}\right]\left[\frac{(x-\mu)^2}{2\sigma^5} - \frac{1}{2\sigma^3}\right]e^{-(x-\mu)^2/2\sigma^2} - \left[\frac{\mu^2}{2\sigma^5} - \frac{1}{2\sigma^3}\right]e^{-\mu^2/2\sigma^2}(x-\mu)\left[e^{-(x-\mu)^2/2} + \sigma^{-3}e^{-(x-\mu)^2/2\sigma^2}\right]\right|$$

$$= \frac{1}{8\pi f(x)f(0)}|A + B + C| \text{ where:}$$

$$A = e^{-\mu^2/2-(x-\mu)^2/2}\left[-\mu\{(x-\mu)^2/2\sigma^5 - 1/2\sigma^3\}\right]$$

$$B = e^{-\mu^2/2\sigma^2-(x-\mu)^2/2}(x-\mu)(-\mu^2/2\sigma^5+1/2\sigma^3)$$

$$C = e^{-\mu^2/2\sigma^2-(x-\mu)^2/2\sigma^2}\{x/2\sigma^6 - x\mu(x-\mu)/2\sigma^8\}$$

Now for $0 < \sigma^2 < \varepsilon < 1$ and $|\mu| < \sigma$ so that $e^{-\mu^2/2} < \sigma^{-1}e^{-\mu^2/2\sigma^2}$

$$\sqrt{|h|} > \frac{|A + B + C|}{2e^{-(x-\mu)^2/2}.2\sigma^{-1}e^{-\mu^2/2\sigma^2}}$$

The likelihood can again be taken as $\sigma^{-1}e^{-\mu^2/2\sigma^2}$.

A: $\int_\eta^\varepsilon d\sigma^2 \int_0^\sigma d\mu\ \sigma^{-1}e^{-\mu^2/2\sigma^2}\ \frac{|A|}{f(x)f(0)} < \infty$ due to the presence of $e^{-(x-\mu)^2/2\sigma^2}$ in the integrand.

B: $\int_\eta^\varepsilon d\sigma^2 \int_0^\sigma d\mu\ \sigma^{-1}e^{-\mu^2/2\sigma^2}\ \frac{|B|}{f(x)f(0)}$

$> k\int_\eta^\varepsilon d\sigma^2 \int_0^\sigma d\mu\ \ e^{-\mu^2/2\sigma^2}\left[-\mu^2/2\sigma^5 + 1/2\sigma^3\right]$ for some constant k

$= k\int_\eta^\varepsilon d\sigma^2\ \sigma^{-3}\left[\ \mu e^{-\mu^2/2\sigma^2}\ \right]_0^\sigma = \int_\eta^\varepsilon \frac{d\sigma^2}{\sigma^2}\ e^{-1/2} \longrightarrow \infty$

C: $\int_\eta^\varepsilon d\sigma^2 \int_0^\sigma d\mu\ \sigma^{-1}e^{-\mu^2/2\sigma^2}\ \frac{|C|}{f(x)f(0)}$ is finite.

So under the empirical metric the likelihood does not integrate.

■

Trying to obtain a measure on the whole parameter space may be expecting too much. A natural choice for a 95% confidence region would be a subset of the parameter space with posterior measure 0.95 and minimum prior measure over all such subsets; a neighbourhood of (0,0) will be included in any such confidence region, unless it is artificially excluded. If we have only one

interest parameter we can use projection method to obtain a likelihood measure on the interest parameter space. We take σ^2 as the interest parameter as there are simpler techniques for estimating μ, the overall mean. For the measure on the fibres σ^2=constant, since no pivotal exists the most convenient choice is the likelihood with respect to the measure derived from the metric. For the Fisher metric this technique still fails.

i) Fisher Metric

To obtain a metric on the interest parameter space, $\mathbb{R}^+$, note that the Fisher metric is independent of μ as μ is a location parameter. Therefore the metric on $\mathbb{R}^+$ is simply g_{22} which is $O(1/\sigma^4)$. The vector field downstairs is obtained by averaging $< \nabla\sigma^2, \nabla l >$ over the fibres σ^2=constant.

$$< \nabla\sigma^2, \nabla l > = \frac{1}{g_{22}} \sum \frac{(-1/2\sigma^3 + (x_i-\mu)^2/2\sigma^5)e^{-(x_i-\mu)^2/2\sigma^2}}{e^{-(x_i-\mu)^2/2} + \sigma^{-1}e^{-(x_i-\mu)^2/2\sigma^2}}$$

Because the metric is independent of μ the probability measure on the fibres is $\frac{\Pi f(x_i)d\mu}{\int \Pi f(x_i)d\mu}$.

Each $f(x_i)$ is the sum of two exponentials, so we can decompose $\Pi f(x_i)$ into a sum over 2^n terms

Now, $\int \Pi f(x_i)d\mu = (8\pi)^{-n/2} \sum \int \sigma^{|s|-n} e^{-\Sigma(x_i-\mu)^2/2 - \Sigma(x_i-\mu)^2/2\sigma^2} d\mu$

where the first sum is over subsets s of $\{0,\ldots,n-1\}$, the second sum is over $i \in s$ and the third is over $i \in s^c$. The main contribution will come when $|s| = n$ or $n-1$ since otherwise the integrand will be uniformly small. If $|s| = n$ or $n-1$ its value is $O(1)$ ie bounded above and below as $\sigma^2 \longrightarrow 0$. Expanding the top

line, $\int \frac{1}{g_{22}} \sum \frac{(-1/2\sigma^3 + (x_j-\mu)^2/2\sigma^5)e^{-(x_j-\mu)^2/2\sigma^2}}{e^{-(x_j-\mu)^2/2} + \sigma^{-1}e^{-(x_j-\mu)^2/2\sigma^2}} \Pi f(x_i)d\mu$

$= \frac{1}{\sqrt{8\pi}} \int \frac{1}{g_{22}} \sum (-1/2\sigma^3+(x_j-\mu)^2/2\sigma^5)e^{-(x_j-\mu)^2/2\sigma^2} \prod_{i\neq j} f(x_i)d\mu$

in the same way, we see that any term with $e^{-(x_j-\mu)/2\sigma^2}e^{-(x_i-\mu)/2\sigma^2}$ will give a negligible contribution as $\sigma^2 \longrightarrow 0$. So the leading term (up to constant multiple) is:

$\sum \frac{1}{g_{22}} \int (-1/2\sigma^3+(x_j-\mu)^2/2\sigma^5)e^{-(x_j-\mu)^2/2\sigma^2}e^{-\Sigma(x_i-\mu)^2/2}d\mu$

[the second sum being over $i\neq j$]

which is $\frac{1}{g_{22}}O(1)$ as $\sigma^2 \longrightarrow 0$. So the vector field downstairs (in these coordinates) is $k\sigma^{-4}$ for some k, while the metric is σ^{-4}, to first order. The function whose gradient under this metric is $k\sigma^{-4}$ is $k\sigma^2$. So the likelihood measure is $e^{k\sigma^2}\sigma^{-2}d\sigma^2$. This is useless as it puts all the mass at 0.

ii) Empirical metric.

Again for convenience assume that there are two observations $x,0$. For small σ^2 and μ close to 0 the leading terms give the matrix for h as $\begin{bmatrix} \mu^2/\sigma^4 + x^2 & -\mu(\mu^2-\sigma^2)/2\sigma^6 \\ -\mu(\mu^2-\sigma^2)/2\sigma^6 & (\mu^2-\sigma^2)^2/4\sigma^8 \end{bmatrix}$ to leading order. A possible problem is that this is singular at $\mu=\sigma$. This is inevitable for two observations, though in this case the matrix is very nearly singular for any number of observations. Since the metric depends on both μ and σ^2, the metrics g and $\bar{g}$ will be different. Recall that the essential part of the likelihood is $\sigma^{-1}e^{-\mu^2/2\sigma^2}$ so that the $\frac{\partial}{\partial\sigma^2}$log-likelihood is $-1/2\sigma^2 + \mu^2/2\sigma^4$. This has to be averaged over the fibre of constant σ^2. So $\frac{\partial}{\partial\sigma^2}$log-likelihood downstairs is $O(\sigma^{-2})$, and the likelihood cannot increase faster than a power of σ^{-2} as $\sigma^2 \longrightarrow 0$. The likelihood

will integrate provided the measure downstairs is small enough. Now $< \nabla\sigma^2, \nabla\sigma^2 > = \mu^2\sigma^4(\mu^2-\sigma^2)^{-2}$ to first order. Averaging this quantity over the fibre with respect to likelihood measure yields an infinite value. This would make the metric dual downstairs infinite and the metric singular. Now for more than two observations the metric is never quite singular but the correction involves considering terms such as $e^{-x^2/2\sigma^2}$ where x is a constant observation and σ^2 is small. This ensures that the metric downstairs is decreasing faster than any power of σ^2. The simulations below confirm that no credence is given to small values of σ^2. This behaviour gives further motivation for the choice of harmonic mean metric: it is the places on the manifold where the metric is smallest which represent areas where there is least information; when integrating over a fibre, it is the contribution from these areas which should have greatest impact on the amount of information available at that value of the interest parameter.

■

These methods are fairly easy to implement numerically and the results agree with the above theory in that spikes occur at $\sigma^2 = 0$ in all cases except the last. The empirical metric is easier to calculate as it does not require a separate numerical integration. FIGURE 4.3 and FIGURE 4.4 show the derived likelihood measures for 10 and 50 observations from the projection method with empirical metric. In other cases the equivalent diagram has a spike appearing as $\sigma^2 \longrightarrow 0$.

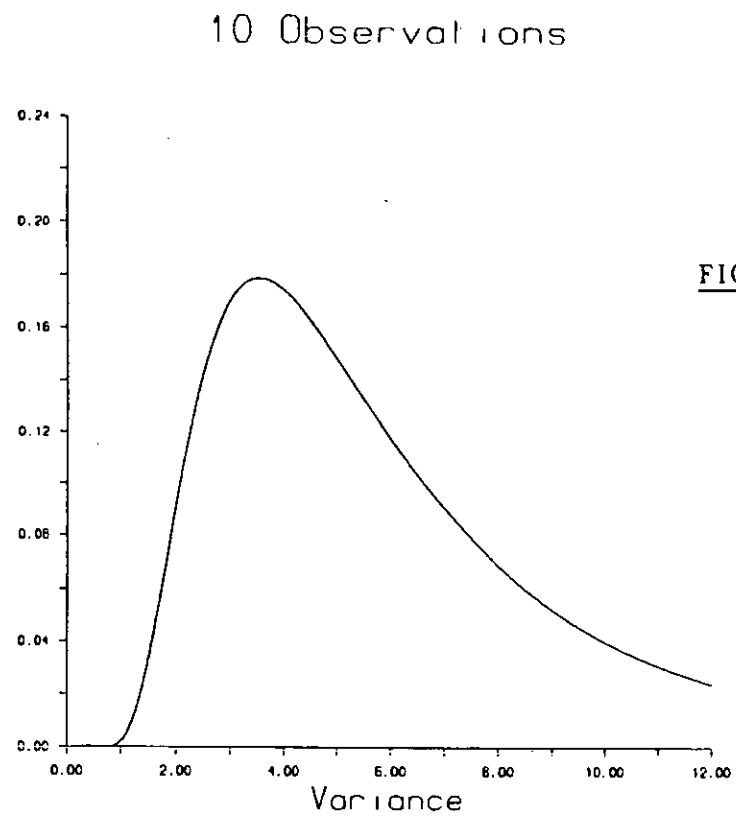


FIGURE 4.3

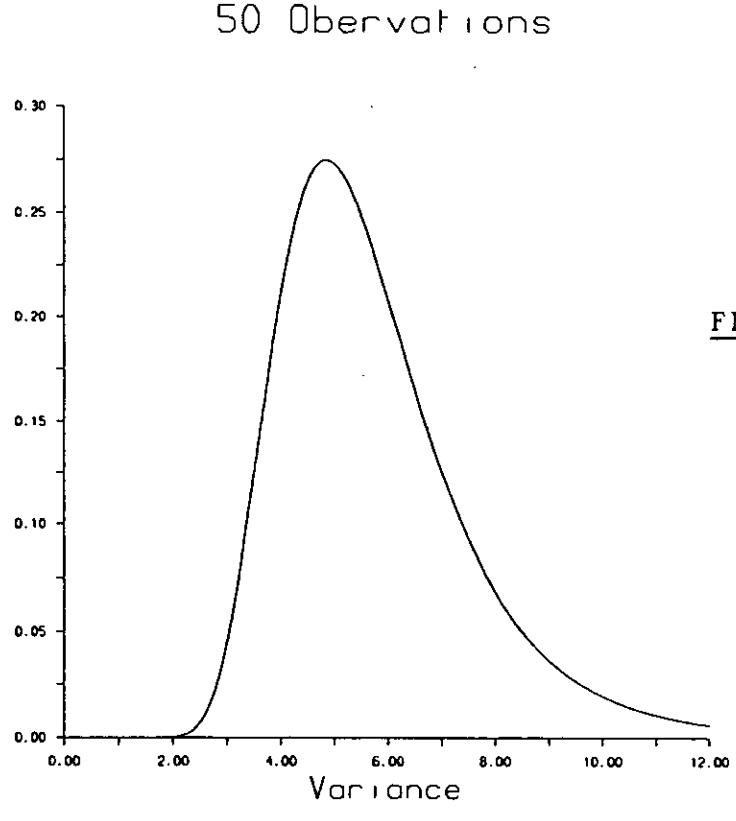


FIGURE 4.4

To summarize: in general cases where asymptotic results do not apply, deriving a metric and a vector field on the interest parameter space is a natural way to proceed as they can directly relate to beliefs about the interest parameters. It is essential to choose a metric that will yield sensible results. In some cases, a metric under which fibres of constant interest parameter are parallel is a good choice. Such a metric can be derived from a general one by averaging over fibres, ie deriving $\bar{g}$ from g.

Chapter 5: Inference from Parameter-Dependent Stochastic Processes

In this chapter we look at a family of cases where the likelihood function itself is not available. These stem from the 1-dimensional diffusion governed by the Ito Equation

$$dX_t = \mu(X_t,\theta,t)dt + \sigma(X_t,\theta,t)dB_t$$

where B_t is a Wiener Process (Brownian Motion). The parameter θ is fixed but unknown and the object is its inference from the observations. We will assume sufficient smoothness.

If the whole path is observable and σ is independent of θ, then the log-likelihood function is given by:

$$l(\theta) = \int \frac{\mu_t(\theta)}{{\sigma_t}^2}\, dX_t - \frac{1}{2}\int \frac{{\mu_t}^2(\theta)}{{\sigma_t}^2}\, dt.$$

and the problem can be tackled by standard techniques.

If σ depends non-trivially on θ then the value of θ can be deduced immediately from the path viz:

$$\sigma(0,\theta,0) = \overline{\underset{t\to 0}{\text{Lim}}}\ \frac{X_t}{\sqrt{(2t\text{loglog}[1/t])}} \text{ a.s.}$$

This is because $\frac{\mu_t}{\sqrt{(2t\text{loglog}[1/t])}} \longrightarrow 0$ and $\overline{\underset{t\to 0}{\text{Lim}}}\ \frac{B_t}{\sqrt{(2t\text{loglog}[1/t])}} = 1$ by the low of the iterated logarithm.

Some interesting cases occur when only partial observations are available. Genon-Catalot and Laredo[9] consider the case $\mu > 0$ and no explicit t-dependence. Only the first hitting times process is observable, ie $H_a = \text{Inf}\{t:X_t > a\}$. Again if σ depended non-trivially on θ, it could be deduced immediately so σ independent of θ was assumed. In this case there is no

explicit form for the likelihood function, but a statistic asymptotically equivalent to the mle as $\sigma \longrightarrow 0$ is exhibited in the paper.

Here we look at the case when only the values of X at integer times can be observed. For simplicity we assume that X is a martingale, and that there is no explicit t-dependence so that $dX_t = \sigma(X_t,\theta)dB_t$. Given observations $x_0=0$, $x_1,\ldots,x_n$, the likelihood function can only be calculated by taking an expectation over all possible paths from $x_0 \to x_1 \to \ldots \to x_n$:

$$l(\theta) = \sum \log \mathbb{E}\left[(\sigma_i^2)^{-1/2}\ e^{-(x_i-x_{i-1})^2/2\sigma_i^2}\ \middle|\ x_{i-1}\right]$$

where $\sigma_i^2 = \int_{i-1}^{i} \sigma^2(X_t,\theta)\ dt$.

The easiest way to see this is to consider that the martingale as a time-changed Brownian Motion. Thus X_i is an observation from this Brownian Motion at a random time T whose distribution depends on X_{i-1}. The conditional density function for X_i is therefore

$$\mathbb{E}\left[(2\pi T)^{-1/2}\ e^{-(x_i-x_{i-1})^2/2T}\ \middle|\ x_{i-1}\right]$$

This is not an easy quantity to calculate or even estimate. So instead we will find some alternative ways of estimating θ. These will not be optimum because the likelihood function is theoretically calculable.

Example 1: $dX_t = \theta\sqrt{1+X_t^2}\ dB_t$

A simple approximation would be to consider each $X_i|X_{i-1}$ as $N\left[X_{i-1},\theta^2(1+X_{i-1}^2)\right]$ which assumes that the function σ stays constant on each unit interval. However the estimator $\frac{1}{n}\sum_{i=1}^{n} (x_i-x_{i-1})^2/(1+x_{i-1}^2)$ is a consistent estimator of $e^{\theta^2}-1$.

By Ito's Formula (see Rogers[19] p60) a function $f(X_t,t,\theta)$ will be a martingale iff $\frac{1}{2}\theta^2(1+X_t^2)f''(X_t,t,\theta) + \dot{f}(X_t,t,\theta) = 0$. The following separable solutions exist:

X_t

$(1 + X_t^2)e^{-\theta^2 t}$

$(X_t^3 + X_t)e^{-3\theta^2 t}$

$(5X_t^4 + 6X_t^2 + 1)e^{-6\theta^2 t}$

$(7X_t^5 + 10X_t^3 + 3X_t)e^{-10\theta^2 t}$

$(21X_t^6 + 35X_t^4 + 15X_t^2 + 1)e^{-15\theta^2 t}$

$\vdots$

From this it can be deduced that;

$$\mathbb{E}\left[(1+X_i^2)/(1+X_{i-1}^2)\,|X_{i-1}\right] = (1+X_{i-1}^2)^{-1}\mathbb{E}\left[(1+X_i^2)e^{-i\theta^2}\,|X_{i-1}\right]e^{i\theta^2}$$

$$= (1+X_{i-1}^2)^{-1}\left[(1+X_{i-1}^2)e^{-(i-1)\theta^2}\right]e^{i\theta^2} \text{ (Martingale Property)}$$

$$= e^{\theta^2} \text{ and}$$

$$\mathbb{V}\left[(1+X_i^2)/(1+X_{i-1}^2)\,|X_{i-1}\right] = \left[(5x_{i-1}^2+1)e^{6\theta^2} + 4e^{\theta^2}\right] \Big/ \left[5x_{i-1}^2+1\right]$$

Since this variance is bounded uniformly in x_{i-1} we have that $\frac{1}{n}\sum_{i=1}^{n} (1+x_i^2)/(1+x_{i-1}^2)$ is a consistent estimator of e^{θ^2}.

To deduce asymptotic normality we apply Theorem 3.2 of Hall and Heyde[11] (p58). To match their notation let

$$X_{ni} = \frac{1}{\sqrt{n}}\left[\frac{1+x_i{}^2}{1+x_{i-1}{}^2} - e^{\theta^2}\right]$$

Lemma:

$\sum_{i=1}^{n} X_{ni}$ tends stably in distribution to a $N(0, e^{6\theta^2} - e^{2\theta^2})$ random variable.

The definition of stability (see Hall[11] p56) is: Y_n converge to Y in distribution on Probability space $(\Omega,\mathcal{F},\mathbb{P})$ and $\forall$ events $E \in \mathcal{F}$ Lim $\mathbb{P}[\{Y_n < y\} \cap E] = Q_y(E)$ exists and tends to $\mathbb{P}(E)$ as $y \longrightarrow \infty$.

Proof:

First note that for any positive random variable Y if $\zeta > 0$ then:

$$\mathbb{E}\left[Y\ I_{\{Y > \zeta\}}\right] = \int_\zeta^\infty \mathbb{P}[Y > x]dx = \int_\zeta^\infty \mathbb{P}[Y^2 > x^2]dx$$

$$= \int_{\zeta^2}^\infty \mathbb{P}[Y^2 > u]du/2\sqrt{u} < \int_{\zeta^2}^\infty \mathbb{P}[Y^2 > u]du/2\zeta = \mathbb{E}\left[Y^2\ I_{\{Y > \zeta\}}\right]\frac{1}{2\zeta}.$$

We use this to check the Lindeberg condition (Hall[11] Cor 3.1).

Let $Y = nX_{ni}^2$. The σ-algebra generated by $X_{n1},\ldots,X_{ni}$ is $\mathcal{F}_{n,i}$.

$$\sum_{i=1}^n \mathbb{E}\left[X_{ni}^2 I_{\{X_{ni}>\epsilon\}}\middle|\mathcal{F}_{n,i-1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y\ I_{\{Y > n\epsilon^2\}}\middle|\mathcal{F}_{n,i-1}\right]$$

$$< \frac{i}{2n^2}\sum_{i=1}^n \mathbb{E}\left[Y^2\ I_{\{Y > n\epsilon^2\}}\middle|\mathcal{F}_{n,i-1}\right] < \frac{1}{2n^2}\sum_{i=1}^n c$$

for some constant c, since $\mathbb{E}\left[\left\{\frac{1+X_i^2}{1+X_{i-1}^2} - e^{\theta^2}\right\}^4\middle|X_{i-1}\right]$ is bounded uniformly in X_{i-1} (this expectation can be calculated explicitly using the martingales exhibited earlier), so $\sum_{i=1}^n \mathbb{E}\left[X_{ni}^2 I_{\{X_{ni} > \epsilon\}}\middle|\mathcal{F}_{n,i-1}\right] \longrightarrow 0$ as $n \longrightarrow \infty$.

We also need $\sum_{i=1}^n \mathbb{E}\left[X_{ni}^2\middle|\mathcal{F}_{n,i-1}\right] \longrightarrow \eta^2$ in probability. To see this consider the process with reflecting boundary at $\pm M$. The stationary measure (see Chapter 2) is $\frac{1}{\theta\sqrt{(1+x^2)}}\,dx$ normalised to integrate to 1 between $\pm M$. By the ergodic theorem for any κ and $\delta > 0$ there will exist an M such that Lim $\frac{1}{n}\left[\Sigma I_{\{X_i < \kappa\}}\right] \longrightarrow \delta$ in probability where M is such that

$$\int_{-\kappa}^{\kappa} \frac{1}{\theta\sqrt{(1+x^2)}}\,dx \Big/ \int_{-M}^{M} \frac{1}{\theta\sqrt{(1+x^2)}}\,dx = \delta.$$

Now $n\mathbb{E}\left[X_{ni}^2 \middle| \mathcal{F}_{n,i-1}\right] = \mathbb{E}\left[\frac{X_i^4+2X_i^2+1}{(X_{i-1}^2+1)^2} \middle| \mathcal{F}_{n,i-1}\right] - e^{2\theta^2}$

$= \mathbb{E}\left[\frac{5X_i^4+6X_i^2+1+4X_i^2+4}{5(X_{i-1}^2+1)^2} \middle| \mathcal{F}_{n,i-1}\right] - e^{2\theta^2}$

$= \left[(5X_{i-1}^2+1)e^{6\theta^2} + 4e^{\theta^2}\right] \Big/ \left[5(X_{i-1}^2+1)\right] - e^{2\theta^2} \longrightarrow e^{6\theta^2} - e^{2\theta^2}$ as $X_{i-1} \longrightarrow \infty$. Since the unrestrained process is stochastically greater than the reflecting process, a standard limit argument will give $\sum_{i=1}^{n} \mathbb{E}\left[X_{ni}^2 \middle| \mathcal{F}_{n,i-1}\right] \longrightarrow e^{6\theta^2} - e^{2\theta^2}$ in probability.

■

The convergence may be too slow to be useful in practice: on a simulation of 500 observations with $\theta=1$ the sample mean of the $|X_i|$'s was ~ 2.2 whereas for the convergence we need the $|X_i|$ to be large most of the time.

▨

Since we can calculate the moments of the martingales given earlier we will look for an estimating procedure based on those martingales. It is clear that $\frac{1}{n}\sum \frac{1+X_i^2}{1+X_{i-1}^2} + (X_i - X_{i-1})\alpha(X_{i-1})$ will be an unbiased estimate of e^{θ^2} for any function α. To minimise the variance we need to take $\alpha(X_{i-1}) = -\frac{X_{i-1}}{1+X_{i-1}^2} e^{\theta^2}(e^{\theta^2}+1)$. The problem is that θ is unknown. Noting that if α depends on $X_1, \ldots, X_{i-1}$ the statistic is still unbiased, a sequential approach could be employed which would involve using the first i-1 observations to estimate e^{θ^2} and using that to make the ith statistic. This idea fails because the estimate does not have variance which tends to zero. Indeed the series can oscillate wildly.

Instead find the estimate of e^{θ^2} with $\alpha=0$ and then use this estimate in the formula for α to obtain an improved estimate. This final estimate will be consistent since the initial estimate is consistent for e^{θ^2}. However it will not be unbiased. To remove first order bias we can use the Jackknife -see Cox[4] p261. This has the added attraction that there is a formula for estimating the variance of the estimate. To calculate the Jackknife estimate let t_n denote the estimate based on all the observations and $t_{n-1,i}$ denote the estimate based on all the observations except the ith. Let $\overline{t_{n-1}} = \Sigma\, t_{n-1,i}/n$. Then the Jackknife estimate is $nt_n - (n-1)\overline{t_{n-1}}$. Its variance can be estimated by $\frac{n-1}{n}\sum\left[t_{n-1,i} - \overline{t_{n-1}}\right]^2$. The following table gives results of simulations: five paths were simulated up to time t=500 with $\theta=1$. The estimates of $e^{\theta^2} \approx 2.72$ given are: the straight estimate e1 with $\alpha=0$ together with its conditional variance, Var (the expression for Var involves θ so this is estimated by e1); the modified estimate e2 with $\alpha=\alpha(e1)$; the jackknife correction of e2 and the jackknife estimate of variance.

Path	e1	Var	e2	JK	Jvar
1	2.70	0.43	2.57	2.66	0.09
2	3.07	0.96	2.87	3.11	0.45
3	2.17	0.11	2.46	2.48	0.03
4	2.46	0.25	2.46	2.50	0.04
5	2.56	0.32	2.69	2.75	0.04

The values obtained for Var should be compared with $\frac{1}{n}(e^{6\theta^2}-e^{2\theta^2})$ ≈ 0.79. Certainly incorporating the (X_i-X_{i-1}) term is to be recommended. It is surprising how much structural difference there is between the simulations, in particular the wide range of values for Var and Jvar. One reason for this is that the distribution for X_i^2 is very skew and the values obtained depend significantly on a few very large observations. If we see too few large observations we get an under-estimate and a low report of variance; if we see too many large observations we get an over estimate and a high report of variance. For this reason it is unsatisfactory to simply assume that JK has a normal distribution with mean the true value and variance Jvar. A better way to express our belief in possible parameter values is to use pseudo-likelihoods, which will be introduced via *example 2*.

Example 2: $dX_t = \sqrt{\theta^2+X_t^2}\, dB_t$

The martingales for this example are closely related to those in the previous example:

X_t

$(\theta^2 + X_t^2)e^{-t}$

$(X_t^3 + \theta^2 X_t)e^{-3t}$

$(5X_t^4 + 6\theta^2 X_t^2 + \theta^4)e^{-6t}$

$(7X_t^5 + 10\theta^2 X_t^3 + 3\theta^4 X_t)e^{-10t}$

$(21X_t^6 + 35\theta^2 X_t^4 + 15\theta^4 X_t^2 + \theta^6)e^{-15t}$

$\vdots$

As before there is no proper stationary distribution. However the problem here is that the obvious statistics to use are $(X_i^2 - eX_{i-1}^2)/(e-1)$

which has expectation θ^2 but conditional variance

$$\left[e^6X_{i-1}^4 + \frac{6\theta^2}{5}X_{i-1}^2(e^6-e) + \frac{\theta^4}{5}(e^6-e+1)\right] \Big/ (e-1)^2$$
$$- 2e\theta^2X_{i-1}^2/(e-1) - e^2X_{i-1}^4/(e-1)^2 - \theta^4$$

This tends to ∞ as $X_{i-1} \longrightarrow \infty$ so equal weighting will not work: $\mathbb{V}\left[\frac{1}{n(e-1)}\sum X_i^2 - eX_{i-1}^2\right] \longrightarrow \infty$. In fact it will be impossible to get any estimate whose variance is $O(1/n)$. It will be necessary to weight the terms according to their variance. A convenient way to do this is to follow the theory of optimum estimation in Godambe[10]. The set-up here is to restrict consideration to real functions $h_i(X_1,\ldots,X_i,\theta)$ such that $\mathbb{E}[h_i|\mathcal{F}_{i-1}] = 0$ and then find functions $a_i(X_1,\ldots,X_i,\theta)$ to produce an estimating function $g=\Sigma\, h_i a_{i-1}$. It is clear from conditional expectations that $\mathbb{E}[g] = 0$. The optimum such g is defined as the one which minimizes $\mathbb{E}[g^2]/\{\mathbb{E}[\partial g/\partial\theta]\}^2$. It is proved that the optimum functions a to take are $a_{i-1}^* = \mathbb{E}[\partial h_i/\partial\theta|\mathcal{F}_{i-1}]/\mathbb{E}[h_i^2|\mathcal{F}_{i-1}]$. Though not explicitly stated this all easily extends to the case when h and a are vector-valued. By the same proof the vector $a_{i-1}^* = \left[\mathbb{E}[h_i^T h_i|\mathcal{F}_{i-1}]\right]^{-1}\mathbb{E}[\partial h_i/\partial\theta|\mathcal{F}_{i-1}]$. In the example we take

$$h_i = \begin{bmatrix} X_i - X_{i-1} \\ (\theta^2 + X_i^2) - (\theta^2 + X_{i-1}^2)e \\ (X_i^3 + \theta^2X_i) - (X_{i-1}^3 + \theta^2X_{i-1})e^3 \end{bmatrix}$$

Since we are going to find the optimum scaling all we need is a statistic with conditional expectation 0 and it does not matter

at this stage how it is scaled. There is no limit on how many of the martingales we choose to use. but those involving higher powers of X will be less useful as they will have a higher variance.

The estimator is calculated by setting the estimating function equal to 0.

Five simulations were made up to time 500 with $\theta=1$. Estimates were made using the first two martingales, and the first three martingales and the estimate $E = n^{-1}\Sigma (X_i^2 - eX_{i-1}^2)/(e-1)$ is given.

Data Set	1st	2nd	3rd	4th	5th
2 martingales	0.84	0.88	1.01	0.96	0.89
3 martingales	0.85	0.87	1.02	0.99	0.89
E	-267	-25.2	-97.8	-96.7	-39.3

Considering three martingales instead of two makes very little difference to the estimates. An upper 5% confidence limit for θ for the first data set could be found by simulating many paths for various values of θ and seeing which gave an estimate as low as 0.85 5% of the time. This method is impractically time-consuming so a better approach is to use pseudo-likelihood functions.

Godambe[10] suggests that the function $\mathcal{U} = -\Sigma\, h_i^T a_{i-1}^*$ can be used as a pseudo-score statistic and that $\mathcal{J} = \Sigma\, a_{i-1}^{*T} \mathbb{E}\left[h_i^T h_i \,|\, \mathcal{F}_{i-1}\right]^{-1} a_{i-1}^*$ can be used as a pseudo-observed Fisher Information. There is some theory behind quasi-likelihoods, which are the closest approximation to the

likelihood when only the mean and variance of certain statistics are available - see Wedderburn[22]. The pseudo-likelihood is the analagous construction under the current set-up. The pseudo-score statistic, $\mathcal{S}$ corresponds to $\frac{\partial}{\partial\theta}$[log-likelihood] so to obtain a pseudo-likelihood, $\mathcal{L}$ we have to integrate this up and take the exponential. The pseudo-volume element is given by $\sqrt{\mathcal{J}}\,d\theta$ and this gives us our pseudo-likelihood measure. The five pseudo-likelihood measures for the five simulations are shown in FIGURE 5.2 by plotting $\mathcal{L}\sqrt{\mathcal{J}}$ against θ, and have been normalised to integrate to 1.

Shown in FIGURE 5.1 is the pseudo-likelihood measure for five simulations of length 500 from *example 1*. Likelihood measures are a good way of presenting belief in the different parameter values because:

i) The domain is naturally the whole parameter space: certain estimating methods can produce estimates outside the parameter space which can only be put back artificially, eg $\mu_i \sim N(0,1)$, $X_i \sim N(\mu_i,\sigma^2)$ $\Rightarrow X_i \sim N(0,\sigma^2+1)$. Estimate σ^2 by (sample variance of X)-1.

ii) The natural way of interpreting FIGURE 5.1 is to compare two regions of the parameter space by comparing the areas under the graph. It can only be valid to do this if the area under the graph is the same as the area under the likelihood function when integrated with respect to some volume element on the parameter space.

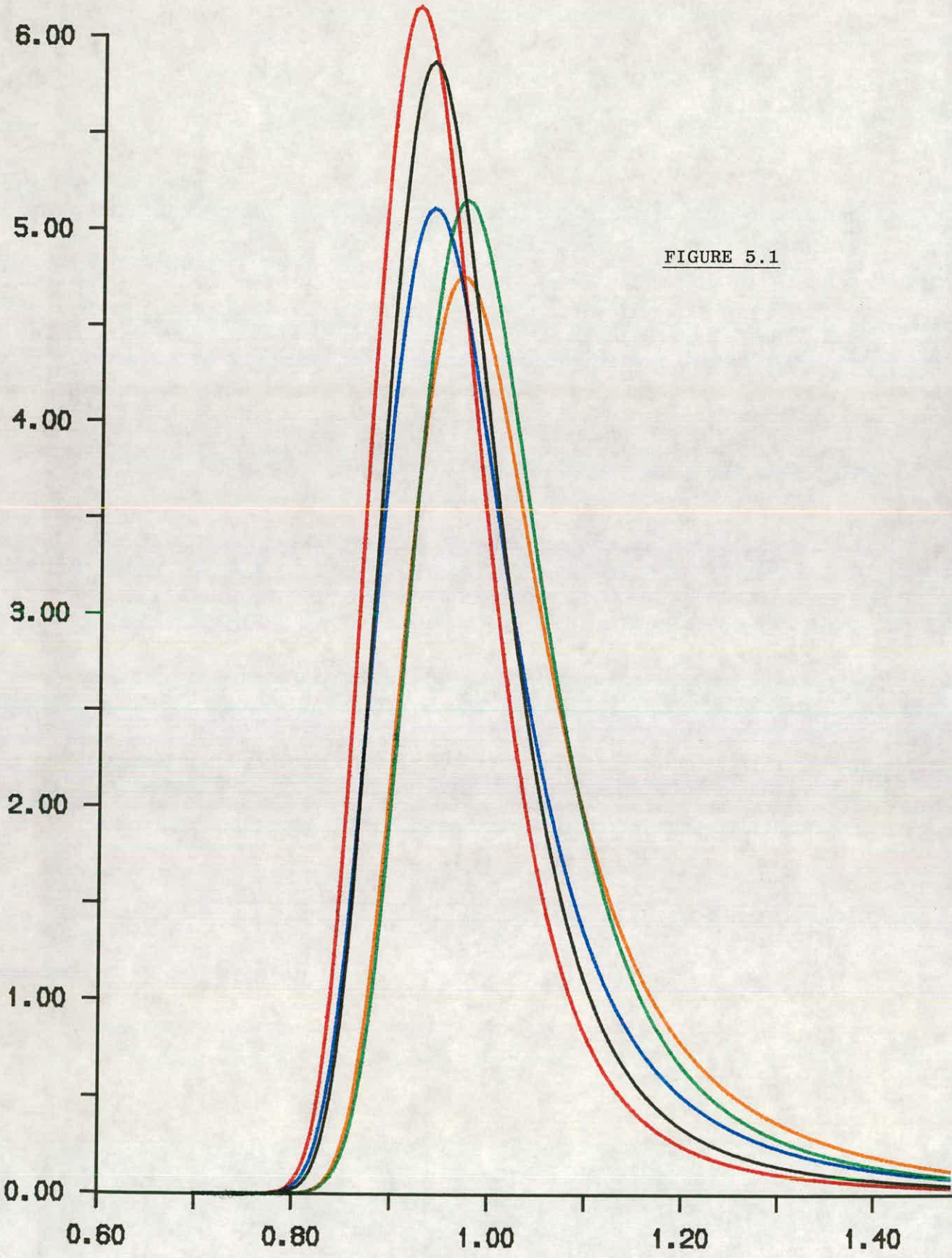


FIGURE 5.1

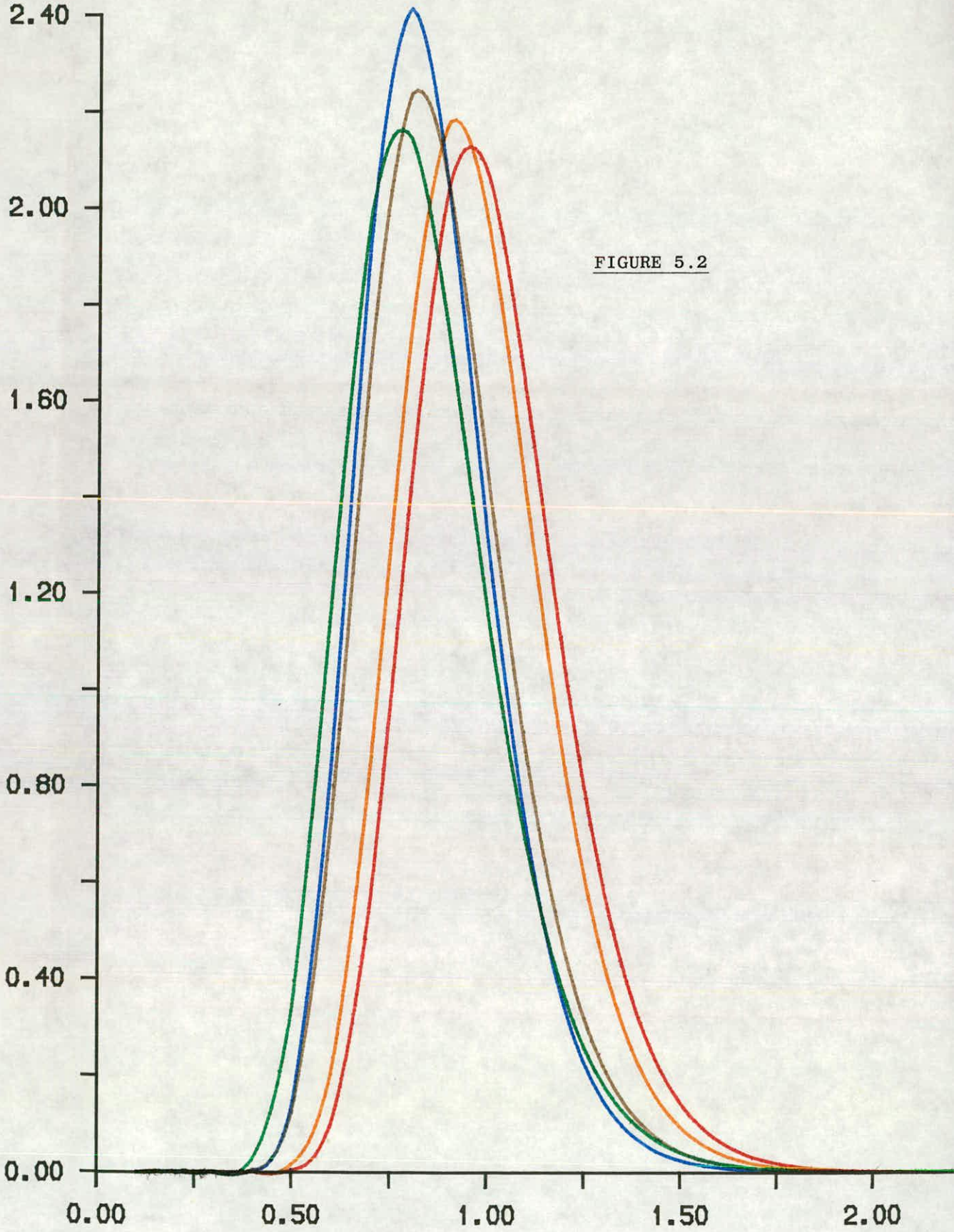


FIGURE 5.2

Chapter 6 Randomly Started Signals with White Noise

This chapter studies the problem of signal detection when there is added white noise. The problem was formulated and explored in Davis (1984)[6]. The approach here will be different so the problem will be restated but keeping notation consistent as far as possible.

Let $\theta \sim U(0,1)$ and B_t an independent Brownian motion, $0 \leqslant t \leqslant 2$.
Let ε be a known constant.
Let $X_t = B_t + \varepsilon\sqrt{(t-\theta)^+} \wedge 1$, where $a \wedge b = \text{Min } \{a,b\}$.

$$(t-\theta)^+ = \text{Max } \{t-\theta, 0\}.$$

Let ξ be the posterior distribution for θ based on observation of $\{X_t : 0 \leqslant t \leqslant 2\}$.

Theorem:

I. If $\varepsilon > \sqrt{8}$ then ξ is (a.s.) a delta function at the true value of θ.

II. If $\varepsilon < \sqrt{8}$ then if $A \subset (0,1)$ has positive (Lebesgue) measure then $\xi(A) > 0$ a.s.

An example of another type of change-point problem is when the Brownian path has a constant drift starting at an unknown point. If the whole path is observed up to time ∞, we will be able to tell almost surely whether there was a drift and what its size was - see for example Pollak[18]. However it will be impossible to pinpoint exactly where the drift started. In the current problem, the drift is so steep at its moment of arrival that (for $\varepsilon > \sqrt{8}$) this time can be pinpointed exactly. The signal

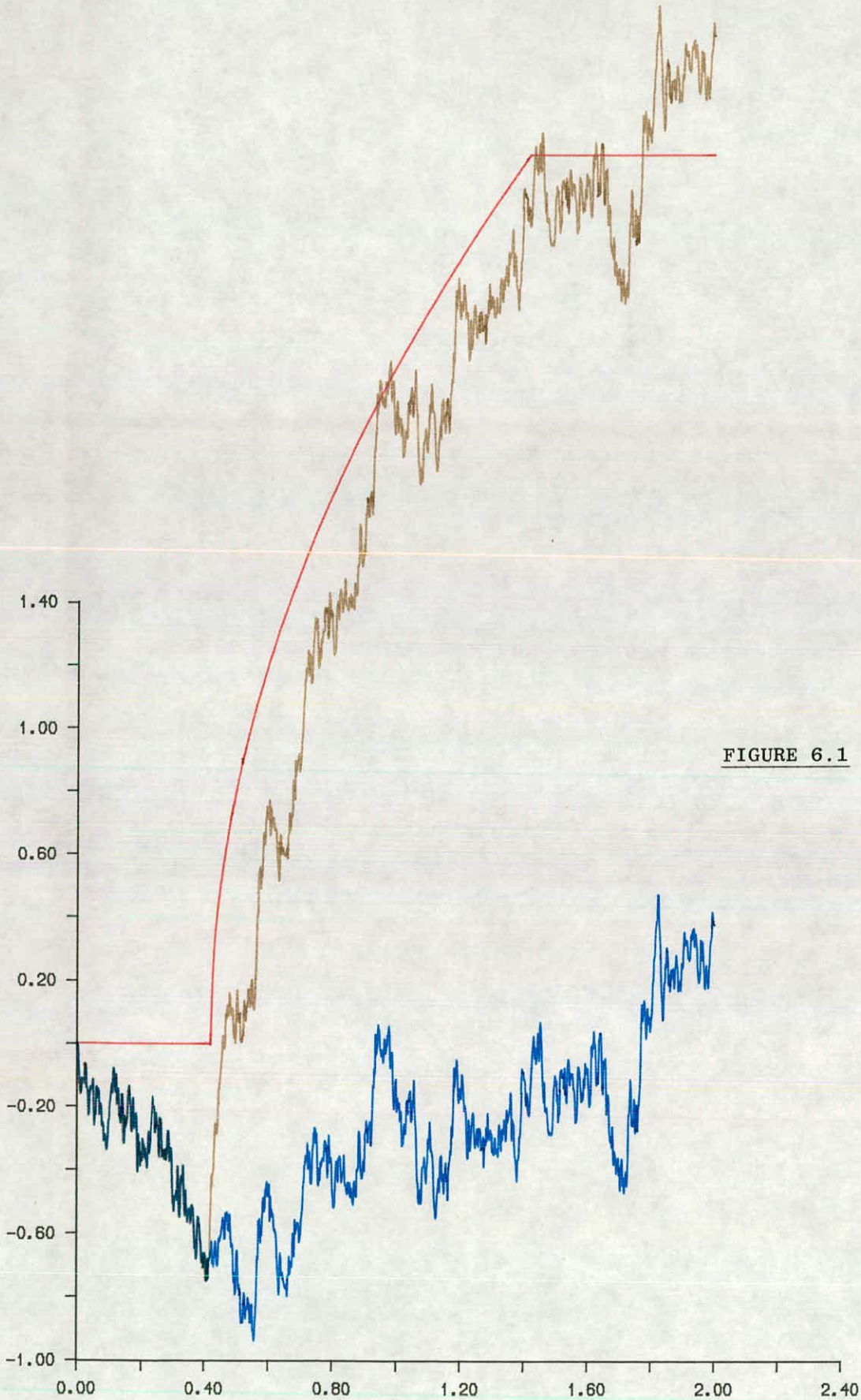


FIGURE 6.1

under consideration here is made to last exactly 1 unit of time for convenience so that stationarity arguments can be used. However for any δ, all the useful information arrives in the first δ seconds after signal arrival so the formulation is essentially equivalent to that of Davis. Figure 6.1 shows a typical Brownian Path in blue, the signal function in red and the sum of these, which is what is observed, in brown.

Davis compares Wiener measure μ, and the measure γ_ε on the space of paths given that a signal arrived at some random time. This measure γ_ε is not simple to understand: if a set of continuous paths, S, on $[0,2]$ has measure $\gamma_\varepsilon(S)$ then if θ is selected from $U(0,1)$ and the path $X_t = B_t + \varepsilon\sqrt{(t-\theta)^+}\wedge 1$ is generated then $\mathbb{P}[\{X_t\} \in S] = \gamma_\varepsilon(S)$.

Davis proves that for $\varepsilon > \sqrt{8}$, μ and γ_ε are absolutely singular. This means that there exists a set of paths S with $\mu(S) = 1$ and $\gamma_\varepsilon(S) = 0$. So if we were told that a certain path were a brownian motion which might or might not have a signal added at a random time, then we could tell almost surely if there were a signal there or not. The theorem above seems to go further in that it shows how to detect the signal arrival time by looking at the posterior distribution.

Lemma: If the measures are singular then given a path which is known to contain a signal, the arrival time can (almost surely) be detected.

Proof: Decompose (0,1) into intervals of length δ and for each interval check whether a signal arrived. The signal actually arrived in exactly one of these intervals so we can (almost surely) decide which one. This narrows the signal arrival time down to an interval of arbitary width δ. ∎

Davis' result for $\varepsilon < 2$ is that μ and γ_ε are absolutely continuous; this is proved using L^2 convergence arguments.

The result above for $\varepsilon < \sqrt{8}$ implies absolute continuity of the two measures. Suppose there were a set of paths S with positive Wiener measure but zero γ_ε-measure. Given a path which is known to contain a signal with random starting point, there is a positive probability that the first part of the path, $X_t : t < 1/2$ matches the first part of a path in S. Since $\gamma_\varepsilon(S) = 0$, we know that a signal cannot have arrived in (0,1/2) and this contradicts the theorem.

The problem of detecting θ is related to the study of Brownian Fast Points - see Davis (1985)[7]. Given a fixed time τ, we can decide from the path whether a signal arrived at time τ no matter how small ε is. Simply note that $(X_{\tau+2^{-k}} - X_{\tau+2^{-k-1}})2^{k/2}$ is an independent Gaussian sequence with variance 1 and mean $\varepsilon\left[2^{-k/2} - 2^{-(k+1)/2}\right]2^{k/2} = \varepsilon(1-1/\sqrt{2})$ if a signal arrives at time τ. Then $\frac{1}{n}\sum_{k=1}^{n} (X_{\tau+2^{-k}} - X_{\tau+2^{-k-1}})2^{k/2} \longrightarrow \varepsilon(1-1/\sqrt{2})$ by the strong law of large numbers, or to 0 if there is no signal at τ. This argument is not directly relevant here since there are an uncountable number of points at which the signal could arrive.

There will be points in (0,1) where the path suddenly increases at an unusually fast rate - so-called fast points, such as the last exit time from 0. It is these points which could be mistaken for the signal if ε is too small.

The approach used here is to consider U(0,1) as a prior distribution for the signal arrival and then calculate the posterior given the path. Thus we consider the process to be a parameter-dependent stochastic process with parameter θ, the arrival time of the signal. In well-behaved parameter-dependent stochastic processes, the Ito Calculus gives a direct expression for the log likelihood function (see §5). In this case because the drift is not a smooth function of the parameter, the log-likelihood thus obtained is not continuous. At the true value it is infinite a.s. At any predetermined value of θ it is $-\infty$ a.s. However it is not $-\infty$ everywhere as there are an uncountable number of possible values for θ. It is true however that the log likelihood function is a.s. $-\infty$ at every rational point and therefore that the posterior for θ is supported on a totally disconnected set.

The method used here is to restrict observations to $\{X_{k.2^{-n}}\}$. Based on just a finite number of observations we are bound to get a continuous posterior density, ξ_n. Convergence arguments can be used to show that as $n \longrightarrow \infty$, $\xi_n d\theta$ will converge weakly to the posterior measure given the whole path. [Recall that weak convergence means that the integral over any fixed interval converges.]

The plan of the proof is to show that the $\log\xi_n$ take the form of a stationary zero-mean Gaussian process plus a positive deterministic part centred at the true θ. *Lemma 6.3* will be used to show that for $\varepsilon > \sqrt{8}$, the maximum of the Gaussian Process away from the true θ is lower than the value of the deterministic part at the true θ, and so the maximum likelihood estimator will almost surely be consistent.

For $\varepsilon < \sqrt{8}$, we show that with positive probability, the log-likelihood function ξ_n attains its maximum at $\hat{\theta}_n$ outside a fixed neighbourhood of θ_0 (the true value), for sufficiently large n. We then compare equal neighbourhoods of θ and $\hat{\theta}$ and use stationarity arguments to show that the values of ξ_n in a neighbourhood of $\hat{\theta}$ are greater than those in a neighbourhood of θ. This will show that the area under the posterior in a neighbourhood of $\hat{\theta}$ is greater than the corresponding area in a neighbourhood of θ with positive probability. This occurs for sufficiently large n and is incompatible with weak convergence to a delta function at θ.

Dealing with the Stationary Gaussian Process presented an interesting problem because for each n the sequence was of fixed length, but for each new n the covariance structure changed so standard asymptotic results cannot be applied. The main asymptotic result used is given in *lemma 6.3*. The preparatory lemmas needed will now be presented.

Lemma 6.1: Let $Q_i \overset{iid}{\sim} N(0,1)$, $i \in \mathbb{N}$. Let $M^Q(2^n) = \text{Max}\{Q_i : i < 2^n\}$. Then

$\forall \epsilon \ \mathbb{P}\left[1-\epsilon < \frac{M^Q(2^n)}{\sqrt{(2n\log 2)}} < 1+\epsilon\right] > 1 - e^{-\epsilon n \log 2}$ for sufficiently large n.

Proof: First we need some well-known inequalities for the normal distribution.

Let $X \sim N(0,1)$. Then $\mathbb{P}[X > m] = \int_m^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

$< \int_m^\infty \frac{x}{m\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{m\sqrt{2\pi}} e^{-m^2/2}$

Let $\frac{\pi}{4} > \eta > 0$. Then $\mathbb{P}[X > m] = \sqrt{\int_m^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \int_m^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy}$

$> \sqrt{\int_{m(\eta+1)\sqrt{2}}^\infty dr \int_{\pi/4-\eta/2}^{\pi/4+\eta/2} r d\theta \frac{1}{2\pi} e^{-r^2/2}} = \sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}$

$= \sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}$

$\mathbb{P}\left[M^Q(2^n) < (1+\epsilon)\sqrt{(2n\log 2)}\right] = \mathbb{P}\left[X < (1+\epsilon)\sqrt{(2n\log 2)}\right]^{2^n}$

$> \left[1 - \frac{1}{m\sqrt{2\pi}} e^{-m^2/2}\right]^{2^n}$ where $m=(1+\epsilon)\sqrt{(2n\log 2)}$

$> 1 - \frac{1}{m\sqrt{2\pi}} e^{-m^2/2} e^{n\log 2} > 1 - e^{-\epsilon n \log 2}/2$ for sufficiently large n.

Let $\eta > 0$ be such that $(1-\epsilon)^2(\eta+1)^2 = 1-\epsilon$

$\mathbb{P}\left[M^Q(2^n) > (1-\epsilon)\sqrt{(2n\log 2)}\right] = 1 - \mathbb{P}\left[X < (1-\epsilon)\sqrt{(2n\log 2)}\right]^{2^n}$

$> 1 - \left[1 - \sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}\right]^{2^n}$ where $m=(1-\epsilon)\sqrt{(2n\log 2)}$

$\left[1 - \sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}\right]^{2^n} = \exp\left[2^n \log\left[1 - \sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}\right]\right]$

$< \exp\left[2^n\left[-\sqrt{\frac{\eta}{2\pi}} e^{-m^2(\eta+1)^2/2}\right]\right] = \exp\left[-\sqrt{\frac{\eta}{2\pi}} e^{\epsilon n \log 2}\right]$

$< \exp[-\epsilon n \log 2]/2$ for sufficiently high n. So

$\mathbb{P}\left[M^Q(2^n) > (1-\epsilon)\sqrt{(2n\log 2)}\right] > 1 - e^{-\epsilon n \log 2}/2$

■

See Galambos[8] for sharper properties of the extreme value.

Lemma 6.2: (D. Slepian):

Let R_i, S_i $:0 \leqslant i < n$ be zero-mean Gaussian processes with variance 1. Let $M^R = \text{Max}\{R_i : i<n\}$, $M^S = \text{Max}\{S_i : i<n\}$.

If $\forall i,j,\ 0 \leqslant \text{Cov}(S_i,S_j) \leqslant \text{Cov}(R_i,R_j)$ then M^S is stochastically greater than M^R.

and more generally $\forall a_i,\ \mathbb{P}\left[\cup \{S_i < a_i\}\right] \leqslant \mathbb{P}\left[\cup \{R_i < a_i\}\right]$

Proof: See Tong[21] p8. ∎

Lemma 6.3: Let $0 < \beta < 1$ and let

$R_n(i) : 0 \leqslant i < \beta 2^n$ be a sequence of stationary zero-mean, unit-variance Gaussian processes (SGP) with $\text{Cov}[R_n(i),R_n(i+r)] = \frac{\log[2^n/(r+1)]}{n\log 2}$ $r \geqslant 0$.

Let $M_n^R = \text{Max}\{R_n(i) : 0 \leqslant i < \beta 2^n\}$

Then $\frac{M_n^R}{\sqrt{(2n\log 2)}} \longrightarrow 1$ in probability.

Proof: Let $\eta > 0$. Let $Q_n(i)$ $0 \leqslant i < \beta 2^n$ be an SGP with zero covariance, ie $Q_n(i) \overset{iid}{\sim} N(0,1)$.

Let $M^Q(\beta 2^n) = \text{Max}\{Q_n(i) : 0 \leqslant i < \beta 2^n\}$.

Then by *lemma 6.1* $\exists N_0$ st $n > N_0 \Rightarrow$

$\mathbb{P}[M^Q(\beta 2^n)/\sqrt{2n\log 2} < 1+\eta] > 1-\eta/2$.

By *lemma 6.2* $\mathbb{P}[M_n^R/\sqrt{(2n\log 2)} < 1+\eta] > \mathbb{P}[M^Q(\beta 2^n)/\sqrt{2n\log 2} < 1+\eta]$

Hence $\mathbb{P}[M_n^R/\sqrt{2n\log 2} < 1+\eta] > 1-\eta/2$. (*)

Let b be such that $\mathbb{P}[N > -b] > 1-\eta/4$ where $N \sim N(0,1)$.

Let k be sufficiently large that

$$\left[\frac{k-2}{k} - \frac{b}{k\sqrt{[2(k-1)\log 2]}}\right]\sqrt{(1-1/k)} > 1-\eta$$

Let $N_1 > k^2$ be such that $n > N_1 \Rightarrow$

$\mathbb{P}\left[\frac{M^Q(\beta 2^{n/k-1})}{\sqrt{[2(n/k-1)\log 2]}} > 1-1/k\right] > 1-\eta/4k$, for Q a zero-covariance SGP.

For such an n, construct k independent SGP's ,

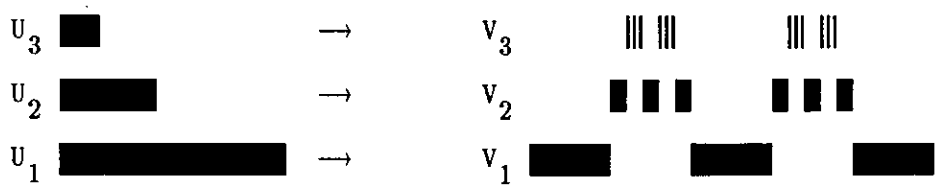
$\{U_j(i): 0 \leqslant i < \beta 2^{n-j}\}$ for $0 \leqslant j < k$, each with variance 1/k, and

$\mathrm{Cov}[U_j(i), U_j(i+r)] = (1-r2^{jn/k-n})^+$ -ie tent-shaped covariance.

Split each sequence into blocks of length $2^{n(1-j/k)}$. Note that the first term in each block is independent of the last term. Now construct non-stationary sequences V_j.

The blocks of the U_1 are copied to the V_1 alternating with gaps of the same length as the blocks $(2^{n(1-1/k)})$.

Then for j=2,...,k-1, the blocks of the U_j are copied to the V_j alternating with gaps of the same length as the blocks, but only above gaps in the V_1 , V_2 ,...,V_{j-1}. The gaps in each V_j are then filled with identical copies of the same term which appears at both ends. Thus each sequence V_j is of length $\beta 2^n$.



$\forall i\ V_0(i) = V_0(0) \sim N(0,1/k)$.

By construction, only one of the V's changes at any time. Let $V=V_0+\ldots+V_{k-1}$. The fastest possible decay of covariance of V is seen to be the straight line segments which dominate $\frac{\log[2^n/(r+1)]}{n\log 2}$.

Consider just those terms in V_1 which come at the end of a block, ie $U_1(0)$, $U_1(2^{n(1-1/k)})$, $U_1(2.2^{n(1-1/k)})$,....

There are $\beta 2^n/(2.2^{n(1-1/k)}) = \beta 2^{n/k-1}$ such terms and they are all independent. Let M_1 be the maximum of these terms.

$$\mathbb{P}\left[\frac{M_1/\sqrt{(1/k)}}{\sqrt{[2(n/k-1)\log 2]}} > 1-1/k\right] > 1-\eta/4k \qquad (\dagger)$$

Now consider those terms in V_2 at the end of blocks which are over the gap filled by M_1. Again there are

$\beta 2^{n(1-1/k)}/(2.2^{n(1-2/k)}) = \beta 2^{n/k-1}$ such values all independent.

So if M_2 is the maximum, $\mathbb{P}\left[\frac{M_2/\sqrt{(1/k)}}{\sqrt{[2(n/k-1)\log 2]}} > 1-1/k\right] > 1-\eta/4k$

Continue finding the maxima in this way. We also have:

$\mathbb{P}(V_0\sqrt{k} > -b) > 1-\eta/4$.

Let $M = V_0+M_1+\ldots+M_{k-1}$. Summing all these inequalities:

$$\mathbb{P}\left[\frac{M\sqrt{k}}{\sqrt{[2(n/k-1)\log 2]}} > k-1 - \frac{k-1}{k} - \frac{b}{\sqrt{[2(n/k-1)\log 2]}}\right] > 1-\eta/2$$

Now $\left[k-1 - \frac{k-1}{k} - \frac{b}{\sqrt{[2(n/k-1)\log 2]}}\right]\frac{\sqrt{[2(n/k-1)\log 2]}}{\sqrt{k}}$

$= \left[\frac{k-2}{k} + \frac{1}{k^2} - \frac{b}{k\sqrt{[2(n/k-1)\log 2]}}\right]\sqrt{1-k/n}\ \sqrt{2n\log 2}$

$\geqslant \left[\frac{k-2}{k} - \frac{b}{k}\right]\sqrt{1-1/k}\ \sqrt{2n\log 2} \geqslant (1-\eta)\sqrt{2n\log 2}$

So $\mathbb{P}\left[\frac{M}{\sqrt{(2n\log 2)}} > 1-\eta\right] > 1-\eta/2$.

Since M is less than $M_V = \mathrm{Max}\{V(i): i < \beta 2^n\}$ and M_V is stochastically less than M_n^R by *lemma 6.2*,

$$\mathbb{P}\left[\frac{M_n^R}{\sqrt{(2n\log 2)}} > 1-\eta\right] > 1-\eta/2.$$

Combine with (*) to give $n > \mathrm{Max}\{N_0, N_1\} \Rightarrow$

$$\mathbb{P}\left[\left|\frac{M_n^R}{\sqrt{(2n\log 2)}} - 1\right| < \eta\right] > 1-\eta$$

Given an η, this holds for sufficiently large n so

$\frac{M_n^R}{\sqrt{(2n\log 2)}} \longrightarrow 1$ in probability ∎

Corollary: We will actually need a stronger result:

Given η, recall $M^R(\beta 2^n/n) = \text{Max}\{R_t : t \leqslant \beta 2^n/n\}$

Then $\mathbb{P}\left[\frac{M^R(\beta 2^n/n)}{\sqrt{(2n\log 2)}} < 1-\eta\right] < \eta/n$ for sufficiently large n. The proof is exactly the same as above. Only searching as far as $\beta 2^n/n$ makes no first-order difference in the size of the maximum. *Lemma 6.3* would give the probability above as less than η; however this can easily be sharpened to η/n because where probabilities are introduced, eg at (†), they can be sharpened using *lemma 6.1*. Note that this corollary is by no means the sharpest possible result.

■

Lemma 6.4: i) $\sum_{i=1}^{2^n} (\sqrt{i} - \sqrt{i-1})^2 = \frac{n}{4}\log 2 + O(1)$

ii) For $h \geqslant 1$,

$$\sum_{i=1}^{2^n} (\sqrt{i} - \sqrt{i-1})(\sqrt{i+h} - \sqrt{i+h-1}) = \frac{1}{4}\log\left[\frac{2^n}{h}\right] + O(1)$$

Proof: i) From standard arguments,

$$\int_{i=1}^{2^n} (\sqrt{i} - \sqrt{i-1})^2 \, di < \sum_{i=1}^{2^n} (\sqrt{i} - \sqrt{i-1})^2$$

$$< 1 + \int_{i=1}^{2^n} (\sqrt{i} - \sqrt{i-1})^2 \, di$$

$\frac{1}{2\sqrt{i}} < (\sqrt{i} - \sqrt{i-1}) < \frac{1}{2\sqrt{(i-1)}}$ from the mean value theorem.

$\int_{i=1}^{2^n} \frac{di}{4i} = \frac{n}{4}\log 2$

ii) Following the same idea as above we need to look at $\int_{i=0}^{2^n} \frac{di}{\sqrt{[i(i+h)]}} = \int_{i=0}^{2^n} \frac{di}{\sqrt{[(i+h/2)^2 - h^2/4]}}$.

$i+h/2 = h/2 \cosh\theta$.

$= \cosh^{-1}[2^{n+1}/h+1] = \log\left[2^n/h\right] + O(1)$

■

Lemma 6.5: Let $R(t):t=1,2,\ldots,M$ be a Stationary Gaussian Process with unit variance and covariance function $\rho(t)=|M-t|\wedge 0$, ie tent-shaped covariance. Then $\max\{R(t):0 \leqslant t \leqslant M\}$ is stochastically less than $2|N(0,1)|$.

Proof: Let $B(t)$ and $B'(t):0 \leqslant t \leqslant M$ be independent Brownian motions. Then $\Lambda(t) = M^{-1/2}[B(t)+B'(M-t)]$ is a (continuous) Stationary Gaussian Process with covariance function $|M-t|\wedge 0$.

$\text{Max}\{B(t):0 \leqslant t \leqslant M\} \sim |N(0,M)|$ and

$\text{Max}\{B'(t):0 \leqslant t \leqslant M\} \sim |N(0,M)|$ independently (see Karatzas[14] p95). So $\text{Max}\{\Lambda(t):0 \leqslant t \leqslant M\}$ is stochastically less than $2M^{-1/2}|N(0,M)| = 2|N(0,1)|$ in distribution, and since $R(t)$ is equal in distribution to Λ at integer points, $\text{Max}\{R(t):0 \leqslant t \leqslant M\}$ is stochastically less than $2|N(0,1)|$.

■

Returning to the original problem recall that $X_t:0 \leqslant t \leqslant 2$ is brownian motion plus a signal arriving at a random time in $(0,1)$.

Let $\mathcal{F}_n = \sigma(X_{2^{-n}}, X_{2.2^{-n}}, X_{3.2^{-n}}, \ldots, X_2)$

$\mathcal{F} = \sigma(X_t:0 \leqslant t \leqslant 2)$, $\mathcal{F}_n \uparrow \mathcal{F}$.

The log-likelihood function for θ based on the values $\{X_{k.2^{-n}}\}$ is:

$$\sum_{k=1}^{2.2^n} -\log(2\pi 2^{-n}) - \frac{1}{2.2^{-n}}\left[\left\{X_{k2^{-n}} - X_{(k-1)2^{-n}}\right\} - \varepsilon\left\{\sqrt{(k2^{-n}-\theta)^+}\wedge 1 - \sqrt{((k-1)2^{-n}-\theta)^+}\wedge 1\right\}\right]^2$$

Up to an additive constant (the last term is constant):

$$l_n(\theta) =$$

$$\sum_{k=[2^n\theta]+1}^{[2^n\theta]+2^n} \epsilon 2^{n/2}\left[X_{k2^{-n}} - X_{(k-1)2^{-n}}\right]\left[\sqrt{k-2^n\theta} - \sqrt{(k-1-2^n\theta)^+}\right]$$

$$- \frac{\epsilon^2}{2}\left[\sqrt{k-2^n\theta} - \sqrt{(k-1-2^n\theta)^+}\right]^2 + \frac{\epsilon^2}{2}\sum_{k=1}^{2^n}(\sqrt{k} - \sqrt{k-1})^2$$

$f_n(\theta)d\theta = \dfrac{e^{l_n(\theta)}d\theta}{\int_0^1 e^{l_n(\theta)}d\theta}$ is the posterior density for θ based on $\{X_{k2^{-n}}\}$.

Lemma 6.6: $f_n d\theta$ converges weakly to the posterior density for θ based on $\{X_t : 0 \leq t \leq 2\}$.

Proof: For Borel set $A \subset [0,2]$, $\int_A f_n(\theta)\, d\theta = \mathbb{P}(\theta \in A|\mathcal{F}_n)$ is a martingale.

So $\int_A f_n(\theta)d\theta \xrightarrow{a.s.} P(\theta \in A|\cup\mathcal{F}_n)$ by martingale convergence. ∎

The essential structure of the log-likelihood function is obtained by looking at its values at the points $\{k.2^{-n}\}$ ie integer values of k. This will be justified when necessary. Suppose $\theta = \theta_0$ is the true value and assume that $\lambda_0 = \theta_0 2^n$ is an integer. Then for $k > \lambda_0$:

$$2^{n/2}\left[X_{k2^{-n}} - X_{(k-1)2^{-n}}\right] = Z_k + \epsilon(\sqrt{k-\lambda_0} - \sqrt{k-\lambda_0-1})$$

For $k < \lambda_0$: $\quad 2^{n/2}\left[X_{k2^{-n}} - X_{(k-1)2^{-n}}\right] = Z_k$

where $Z_k \overset{iid}{\sim} N(0,1)$.

The log-likelihood, $l_n(\theta)$ is split up into a zero-mean Gaussian Stochastic Process, $\epsilon Y_n(\lambda)$ and a deterministic part, $D_n(\lambda)$ for $\lambda = 0,1,\ldots,2^n$ and $\theta = \lambda 2^{-n}$.

$$Y_n(\lambda) = \sum_{k=\lambda+1}^{2^n} Z_k(\sqrt{k-\lambda} - \sqrt{k-\lambda-1}).$$

For $\lambda > \lambda_0$

$$D_n(\lambda) = \varepsilon^2 \sum_{k=\lambda+1}^{\lambda_0+2^n} (\sqrt{k-\lambda_0} - \sqrt{k-\lambda_0-1})(\sqrt{k-\lambda} - \sqrt{k-\lambda-1})$$
$$- \frac{\varepsilon^2}{2} \sum_{k=\lambda+1}^{\lambda+2^n} (\sqrt{k-\lambda} - \sqrt{k-\lambda-1})^2 + \frac{\varepsilon^2}{2} \sum_{k=1}^{2^n} (\sqrt{k} - \sqrt{k-1})^2$$
$$= \varepsilon^2 \sum_{k=\lambda+1}^{\lambda_0+2^n} (\sqrt{k-\lambda_0} - \sqrt{k-\lambda_0-1})(\sqrt{k-\lambda} - \sqrt{k-\lambda-1})$$

For $\lambda < \lambda_0$

$$D_n(\lambda) = \varepsilon^2 \sum_{k=\lambda_0+1}^{\lambda+2^n} (\sqrt{k-\lambda_0} - \sqrt{k-\lambda_0-1})(\sqrt{k-\lambda} - \sqrt{k-\lambda-1})$$
$$- \frac{\varepsilon^2}{2} \sum_{k=\lambda+1}^{\lambda+2^n} (\sqrt{k-\lambda} - \sqrt{k-\lambda-1})^2 + \frac{\varepsilon^2}{2} \sum_{k=1}^{2^n} (\sqrt{k} - \sqrt{k-1})^2$$
$$= \varepsilon^2 \sum_{k=\lambda_0+1}^{\lambda+2^n} (\sqrt{k-\lambda_0} - \sqrt{k-\lambda_0-1})(\sqrt{k-\lambda} - \sqrt{k-\lambda-1})$$

Note that for $\lambda_2 \geq \lambda_1$,

$$\mathbb{C}ov[Y_n(\lambda_1), Y_n(\lambda_2)] = \sum_{k=\lambda_2+1}^{\lambda_1+2^n} (\sqrt{k-\lambda_1} - \sqrt{k-\lambda_1-1})(\sqrt{k-\lambda_2} - \sqrt{k-\lambda_2-1})$$

which has the same form as D_n. Given two zero-mean Gaussian random variables with the same variance we have that

$$\mathbb{E}[Y_n(\lambda_1) | Y_n(\lambda_2)] = Y_n(\lambda_2)\mathbb{C}orr[Y_n(\lambda_1), Y_n(\lambda_2)]$$

Thus $\varepsilon Y_n(\lambda) | Y_n(\lambda_0)=0 + D_n(\lambda)$ is equal in distribution to $\varepsilon Y_n(\lambda) | Y_n(\lambda_0)=D_n(\lambda_0)$ as Gaussian sequences since they have the same expectations and covariance structure.

Approximation of the log-likelihood function:

Let $\delta > 0$. Let $\lambda < 2^n$. From *lemma 6.4*,

$\exists c$ independent of n s.t. $(n\log 2)/4 - c < \mathbb{V}[Y_n(\lambda)] < (n\log 2)/4 + c$

$$1 \leq k \leq 2^n \Rightarrow$$

$$(\log[2^n/k])/4 - c < \mathbb{C}ov[Y_n(\lambda), Y_n(\lambda+k)] < (\log[2^n/k])/4 + c$$

$$|k-\lambda_0| > 2^n\delta \Rightarrow -c < D_n(k) < c$$

$$|k-\lambda_0| < 2^n\delta$$

$$\Rightarrow \frac{\varepsilon^2}{4}(\log[2^n/|k-\lambda_0|]) - c < D_n(k) < \frac{\varepsilon^2}{4}(\log[2^n/|k-\lambda_0|]) + c$$

Proof of I: $\varepsilon > \sqrt{8}$

Let $\delta > 0$. Let $\hat{\lambda}_n$ denote the λ which maximizes $Y_n(\lambda)+D_n(\lambda)$. It is shown that $\hat{\lambda}_n$ lies in $(\lambda_0-2^n\delta,\lambda_0+2^n\delta)$ for sufficiently large n. The true θ may not lie in the set $\{k2^{-n}\}$ but it must be within 2^{-n} of a point in this set so look at $D_n(\lambda_0+1)$. The value $D_n(\lambda_0+1)$ must be less than the deterministic part of l_n at the point in $\{k2^{-n}\}$ closest to the true value.

$$\frac{n\varepsilon^2\log 2}{4} - c < D_n(\lambda_0+1)$$

Let $\alpha=[(\varepsilon-\sqrt{8})\log 2]/8$. *Max* and $\sum$ are taken over $|k-\lambda_0| > 2^n\delta$.

$$\mathbb{P}\left[\textit{Max}\ [\varepsilon Y_n+D_n](k) > [\varepsilon Y_n+D_n](\lambda_0+1)\right]$$

$$< \mathbb{P}[\textit{Max}\ \varepsilon Y_n(k) > \frac{n\varepsilon^2\log 2}{4} - 2c - \alpha n\varepsilon] + \mathbb{P}[\varepsilon Y_n(\lambda_0+1) > \alpha n\varepsilon]$$

$$< \sum \mathbb{P}[Y_n(k) > \frac{n\varepsilon\log 2}{4} - 2c/\varepsilon - \alpha n] + \mathbb{P}[Y_n(\lambda_0+1) > \alpha n]$$

$$< \mathbb{P}\left[Z > \frac{(n\varepsilon\log 2)/4-c-\alpha n}{\sqrt{[(n\log 2)/4+c]}}\right]2^n + \mathbb{P}\left[Z > \frac{\alpha n}{\sqrt{[(n\log 2)/4+c]}}\right]$$

where $Z \sim N(0,1)$.

$$< 2^n\,\frac{1}{2}\exp-\left[\frac{[(n\varepsilon\log 2)/4-c-\alpha n]^2}{2[(n\log 2)/4+c]}\right] + \frac{1}{2}\exp-\left[\frac{\alpha^2n^2}{2[(n\log 2)/4+c]}\right]$$

by well-known inequalities - see *lemma 6.1*.

$$< \exp-\left[n\frac{[(\varepsilon\log 2)/4 - \alpha]^2}{(\log 2)/2} + n\log 2 + c_1\right] + \exp-\left[n\frac{\alpha^2}{(\log 2)/2} + c_1\right] \quad (*)$$

for some c_1 and sufficiently large n.

Now $\frac{[(\varepsilon\log 2)/4 - \alpha]^2}{(\log 2)/2} = 2(1/\sqrt{8} + \varepsilon/8)^2\log 2 > \log 2$

So $(*) < \exp(-c_2n+c_3) + \exp(-c_2n+c_3)$ for some c_2, c_3.

$\sum_n \exp(-c_2n+c_3) + \exp(-c_2n+c_3) < \infty$ so by the Borel-Cantelli Lemma, the event $\{\textit{Max}\ [\varepsilon Y_n+D_n](k) > [\varepsilon Y_n+D_n](\lambda_0+1)\}$ occurs only finitely often. This proves that the path gives enough information to pinpoint the signal arrival time. Therefore the posterior distribution must be a delta function at the true value. ∎

Proof of II: $\varepsilon < \sqrt{8}$:

For the reverse case the first problem is to show that the mle is not almost surely consistent. In fact we show that if $\hat{\lambda}_n$ satisfies $D_n(\hat{\lambda}_n) + \varepsilon Y_n(\hat{\lambda}_n) \geqslant D_n(\lambda) + \varepsilon Y_n(\lambda)\ \forall\lambda$ then $\hat{\lambda}_n \not\longrightarrow \lambda_0$ a.s. To simplify notation we let $\lambda_0=0$, let $\gamma(n)$ be a function satisfying $\gamma(n) \longrightarrow \infty$ and let $\hat{\lambda}^0_n < 2^{n-\gamma(n)}$ satisfy

$$D_n(\hat{\lambda}^0_n) + \varepsilon Y_n(\hat{\lambda}^0_n) \geqslant D_n(\lambda) + \varepsilon Y_n(\lambda)\ \forall\lambda < 2^{n-\gamma(n)}.$$

Thus $\hat{\lambda}^0_n$ is essentially where the log-likelihood takes its maximum in an interval around λ_0 of width $2^{-\gamma(n)}$. We need to show that $D_n(\lambda) + \varepsilon Y_n(\lambda)$ gets at least as high for $\lambda > 2^{n-\gamma(n)}$. Away from λ_0 the contribution from D_n is negligible (but positive). So we need to consider εY_n in the interval $(2^{-\gamma(n)},1)$, but to simplify the proof let Y'_n be equal in distribution and independent of Y and let $\hat{\lambda}^1_n$ satisfy $Y'_n(\hat{\lambda}^1_n) \geqslant Y'_n(\lambda)\ \forall\lambda < 2^n$. What we need to show is that there is a non-zero probability (independent of n) that $(D_n+\varepsilon Y_n)(\hat{\lambda}^0_n) < \varepsilon Y'_n(\hat{\lambda}^1_n)$. If the mle converged almost surely we could find a nest of intervals $I_n = (\lambda_0-2^{-\gamma(n)},\lambda_0+2^{-\gamma(n)})$ with $\mathbb{P}[\hat{\lambda} \in I_n] \longrightarrow 1$.

Firstly we show that because $Y_n/\sqrt{\mathbb{V}[Y_n]}$ is sufficiently close to the SGP in *lemma 6.3*, if $M^Y_n = Max\{Y_n(i):0 \leqslant i < \tilde{\beta}2^n/n\}$ then

$\mathbb{P}\left[\frac{M^Y_n/\sqrt{\mathbb{V}[Y_n]}}{\sqrt{(2n\log 2)}} < 1-\eta\right] < \eta/n$ for sufficiently large n.

Let $h > 4c/\log 2$. Let $Q(i) \overset{iid}{\sim} N(0,1)$.

Let $G_n(i) = Y_n(i2^h) + \sqrt{c}\ Q(i)$, $0 \leqslant i < \tilde{\beta}2^{n-h}$

Then $\mathbb{V}[G_n(i)] > \frac{n}{4}\log 2$ and $Cov[G_n(i),G_n(i+r)] < \frac{1}{4}\log[2^n/(r+1)]$.

So by *lemma 6.2* and *lemma 6.3* with $\beta = 2^{-h}\tilde{\beta}$

$\mathbb{P}\left[\frac{Max\{G_n(i)/\sqrt{\mathbb{V}[G_n(i)]} : i < \beta 2^n/n\}}{\sqrt{(2n\log 2)}} < 1-\eta\right] < \eta/n$ for sufficiently large n.

$\mathbb{V}[G_n(i)]/([n\log 2]/4) \longrightarrow 1$, hence result.

■

This makes it clear that $\varepsilon Y_n'(\hat{\lambda}_n^1)$ is greater than $(D_n+\varepsilon Y_n)(0)$ with probability approaching 1 as $n \longrightarrow \infty$. However this is not strong enough.

Let k satisfy $\frac{\varepsilon}{\sqrt{8}} < k < 1$. Split the domain of Y_n into intervals $I_s = (2^{n-nk^{s-1}}, 2^{n-nk^s})$ where $nk^s > \gamma(n)$ ie $0 < s < \frac{\log n - \log[\gamma(n)]}{\log[1/k]}$. From *lemma 6.4* we have that for $\lambda \in I_s$, $D_n(\lambda)$ is less than $\frac{\varepsilon}{4}nk^{s-1}\log 2 + c$ for some c. Decompose

$$Y_n(\lambda) = \sum_{r=1}^{2^{n(1-k^s)}} Z_{\lambda+r}(\sqrt{r} - \sqrt{r-1})$$

$$+ \sum_{r=2^{n(1-k^s)}+1}^{2^n} Z_{\lambda+r}(\sqrt{r} - \sqrt{r-1}) = T_1(\lambda) + T_2(\lambda).$$

T_1' and T_2' shall refer to the corresponding decomposition of Y_n'. $T_1(\lambda)$ is easy to understand as it has the same behaviour as $Y_n(\lambda)$ with n replaced by $n(1-k^s)$. T_2 should be thought of as a scaled version of T_1. To first order its variance is $\frac{1}{4}\left[\log\left[2^n\right] - \log\left[2^{n(1-k^s)}\right]\right] = \frac{1}{4}nk^s\log 2$.
If we look at every $2^{n(1-k^s)}$th term we find there are 2^{nk^s} terms and the covariance is:

$$\sum_{r=2^{n(1-k^s)}}^{2^n-h2^{n(1-k^s)}} \left[\sqrt{r} - \sqrt{r-1}\right]\left[\sqrt{r+h2^{n(1-k^s)}} - \sqrt{r+h2^{n(1-k^s)}-1}\right]$$

which equals:

$$\frac{1}{4}\left[\log\left[\frac{2^{nk^s}}{h}\right] - \log\left[\frac{2^{n(1-k^s)}+h2^{n(1-k^s)}}{h2^{n(1-k^s)}}\right]\right] = \frac{1}{4}\log\left[\frac{2^{nk^s}}{h}\right] + O(1).$$

So the process $T_2(0), T_2(2^{n(1-k^s)}), T_2(2.2^{n(1-k^s)}), \ldots$ has the same form as Y with n replaced by nk^s so that the process is a scaled version of the parent process.

We also must check that $T_2(\lambda)$ is effectively constant for $\lambda \in \{0,1,2,\ldots,2^{n(1-k^s)}\}$. The covariance function for T_2 is:

$$\rho(h) = \sum_{r=2^{n(1-k^s)}}^{2^n-h} \left[\sqrt{r} - \sqrt{r-1}\right]\left[\sqrt{r+h} - \sqrt{r+h-1}\right] \text{ so}$$

$$\rho(0) - \rho(1) = \left[\sqrt{2^{n(1-k^s)}} - \sqrt{2^{n(1-k^s)}-1}\right]^2 -$$

$$\sum_{2^{n(1-k^s)}}^{2^n} (\sqrt{r} - \sqrt{r-1})(-\sqrt{r+1} + 2\sqrt{r} - \sqrt{r-1}).$$

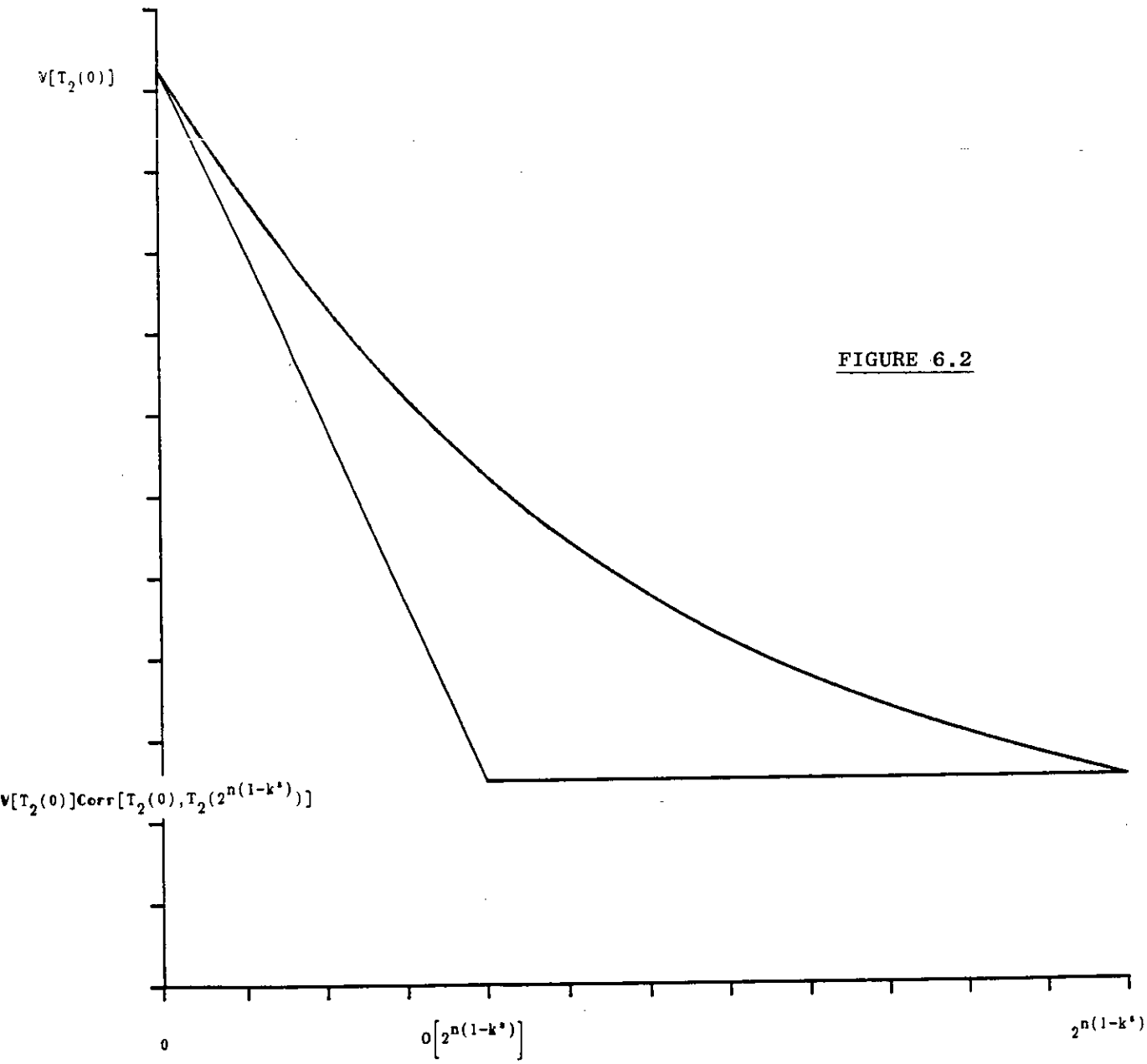


FIGURE 6.2

Since the square root function is concave the sum is positive so that $\rho(0) - \rho(1) < \left[\sqrt{2^{n(1-k^s)}} - \sqrt{2^{n(1-k^s)}-1}\right]^2$
$< \frac{1}{4}\left[2^{n(1-k^s)}-1\right]^{-1}$ by the mean value theorem.
Since the covariance function is convex we can construct a process with a smaller covariance (see FIGURE 6.2) on the set $\{0,1,2,\ldots,2^{n(1-k^s)}\}$ by adding a random variable of variance $\mathbb{V}[T_2(0)]\mathrm{Corr}[T_2(0),T_2(2^{n(1-k^s)})]$ to a suitable process with tent-shaped correlation. This latter process must have variance $\mathbb{V}[T_2(0)]\{1-\mathrm{Corr}[T_2(0),T_2(2^{n(1-k^s)})]\} = O(1)$ and "tent slope" $\frac{1}{4}\left[2^{n(1-k^s)}-1\right]^{-1}$. The length of the tent is therefore $O\left[2^{n(1-k^s)}\right]$ and so the number of tent-lengths up to time $2^{n(1-k^s)}$ is $O(1)$.

By *lemma 6.5* the maximum of this process is $O(1)$ ∎

Recall that the domain of Y_n is a neighbourhood of 0 (the true value) which is shrinking with n. The intervals I_s we consider are given by $s=1,\ldots,\{\log[n]-\log[\gamma(n)]\}/\log[1/k]$ so that $nk^s > \gamma(n)$ is an inequality for s; since we look at every $2^{n(1-k^s)}$th term in T_2, the number of terms we look at is $2^{nk^s} > 2^{\gamma(n)}$ and so for sufficiently large n the number of terms in T_2 will be large enough for the required asymptotic property to hold at all values of s.

We require that given sufficient time the process $T_2'(\lambda)$ will hit the value $\frac{\varepsilon}{4}nk^{s-1}\log 2 + c$. From the choice of k and n we know this will happen in time $\frac{2^n}{nk^s}$ with probability at least $1 - \frac{\eta}{nk^s}$. It makes the argument easier if we start at $\lambda=2^n$ and then search backwards until we find a λ_2 with $T_2'(\lambda_2) > \frac{\varepsilon}{4}nk^{s-1}\log 2 + c$.

Now let $I'_s = \left[\lambda_2-(2^{n-nk^s} - 2^{n-nk^{s-1}}),\lambda_2\right]$, so that I_s and I'_s are the same length. Since I_s is in a fixed position we have that $T_2(2^{n-nk^s}) = O\sqrt{(nk^s)}$. Since T_2 remains constant to first order on this interval, we have that $T'_2 > T_2 + D_n$ when comparing T_2 on I_s and T'_2 on I'_s. Now the idea is to compare T_1 and T'_1 over their respective intervals. Now for $\lambda \in I'_s$, $T'_1(\lambda)$ is independent of everything used so far, because the Z's used in the construction of λ_2 are disjoint from those needed to construct the $T'_1(\lambda):\lambda \in I_s$. This is why λ_2 was found by a backwards search. Hence we must have that $max\{\varepsilon Y'_n(\lambda):\lambda \in I'_n\}$ is stochastically greater than $max\{(D_n+\varepsilon Y_n)(\lambda):\lambda \in I_n\}$.

Before applying this argument, we have to find all the necessary intervals I'_s for $s < (\log n - \log[\gamma(n)])/\log(1/k)$. We hunt for these sequentially. The total time this will take is at most $\sum_{s=0}^{\log n/\log[1/k]} \frac{2^n}{nk^s} = \frac{2^n}{n}\left[(1/k)^{\log n/\log[1/k]}-1\right]/(1/k-1)$ $< 2^n/(1/k-1)$.

The probability of failure is at most $\eta/(1/k-1)$.

Finally we use *lemma 6.2* to compare $max\{(Y_n+D_n)(\lambda):\lambda \in \cup I_s\}$ and $max\{Y'_n(\lambda):\lambda \in \cup I'_s\}$. This shows that there is a probability of at least 1/2 that the maximum occurs away from the true value.

To complete the proof we have to compare areas under the likelihood function. We now know that the true θ and the mle $\hat{\lambda}$ are more than a distance h apart with probability greater than η, for some h,η independent of n. Consider an interval I_h of width h centred on the true θ and an equal interval centred on $\hat{\lambda}$. The values of the likelihood function in the neighbourhood of the $\hat{\lambda}$

are stochastically greater than those in the neighbourhood of the true θ. This is because the stochastic part of the log-likelihood function in a neighbourhood of $\hat{\lambda}$ is simply the Stationary Gaussian Process conditioned on the value at $\hat{\lambda}$ and conditioned to be lower than the value at $\hat{\lambda}$ (at points $\{k2^{-n}\}$). Now because the deterministic part of the log-likelihood function has the same form as the covariance function, the log-likelihood function in a neighbourhood of θ_0 is the Gaussian Process conditioned on the value at the true θ_0 and conditioned to be lower than the value at $\hat{\lambda}$ (at points $\{k2^{-n}\}$). This means that the only difference between the two processes is that the process in a neighbourhood of $\hat{\lambda}$ has a greater deterministic part. So with probability at least 1/2 the area under the likelihood function in a neighbourhood of $\hat{\lambda}$ is greater than the corresponding area in a neighbourhood of θ_0. This occurs for arbitarily large n and is incompatible with the posterior converging to a delta function centred at the true θ.

This proves that the posterior ξ is not almost surely a delta function at the true θ. This is not quite strong enough as it still might be a delta function at θ_0 with probability $p < 1$.

It is (almost surely) impossible that the posterior distribution ξ, for the signal will be a delta function at the wrong value. Recall that $\Lambda_n = \mathbb{P}[\theta \in A | \{X_{k2^{-n}}\}]$ is a martingale. Then $\tilde{\Lambda}_n = \Lambda_{1-1/n}$ is obviously a martingale with $\tilde{\mathcal{F}}_n = \mathcal{F}_{1-1/n}$. Extend $\tilde{\mathcal{F}}$ by making $\tilde{\mathcal{F}}_2 = \Omega$ (ie all information revealed). Then $\tilde{\Lambda}_2 = 1 \iff \theta_0 \in A$. But $\tilde{\Lambda}_1 = 0 \Rightarrow \tilde{\Lambda}_2 = 0$ (as) by the martingale property as $\tilde{\Lambda} \geq 0$.

The proof of the theorem showed that given a small neighbourhood A of θ_0, the area under the likelihood function above A^c was stochastically greater than the area above A. If there were a positive probability that $\xi(A) = 1$ then there must be a positive probability that $\xi(A^c) = 1$ which is impossible from the above argument. Therefore ξ is absolutely continuous with respect to Lebesgue measure on the line, which means that given any interval the posterior probability that θ lies in that interval is strictly positive.

REFERENCES:

[1] Amari,S. (1985): Differential Geometrical Methods in Statistics. *Springer Lecture Notes in Statistics 28*

[2] Berger,J.O. and Wolpert,R.L. (1984): The Likelihood Principle. *Institute of Mathematical Statistics Lecture Notes-Monograph Series Volume 6.*

[3] Carverhill,A.P. , Chappell,M.J. and Elworthy,K.D. (1984): Characteristic Exponents for Stochastic Flows. *Springer Lecture notes in Mathematics vol 1158 "Stochastic Processes - Mathematics and Physics Bielefeld September 1984"* p52-80.

[4] Cox,D.R. and Hinkley,D.V. (1974): Theoretical Statistics. *Chapman and Hall*

[5] Cox,D.R. and Reid,N. (1987): Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society Series B* **49** p1-39

[6] Davis,B. and Monroe,I. (1984): Randomly started signals with White Noise. *The Annals of Probability* **12** p922-925

[7] Davis,B. and Perkins,E. (1985): Brownian Slow Points: the critical case. *The Annals of Probability* **13** p779-803

[8] Galambos,J. (1978): The Asymptotic Theory of Extreme Order Statistics. *Wiley.*

[9] Genon-Catalot,V. and Laredo,C. (1987): Limit Theorems for the first hitting times process of a diffusion and Statistical Applications. *Scandinavian Journal of Statistics* **14** p143

[10] Godambe,V.P. (1985): The Foundations of Finite Sample Estimation in Stochastic Processes. *Biometrika* **72** p419-428

[11] Hall,P. and Heyde,C.C. (1980): Martingale Limit Theory and its Application. *Academic Press.*

[12] Hicks,N.J. (1965): Notes on Differential Geometry. *Van Nostrand Mathematical Studies.*

[13] Kalbfleisch,J.D. and Sprott,D.A. (1970): Application of Likelihood Methods to Models involving Large Numbers of Parameters. *Journal of the Royal Statistical Society Series B* **32** p175-192.

[14] Karatzas,I. and Shreve,S.E. (1988): Brownian Motion and Stochastic Calculus. *Springer-Verlag.*

[15] Kiefer,J. and Wolfowitz,J. (1956): Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27** p887-906

[16] Maritz,J.S. and Lwin,T. (1989): Empirical Bayes Methods. *Chapman and Hall.*

[17] Neymann,J. and Scott,E.L. (1948): Consistent estimates based on partially consistent observations. *Econometrica* **16** p1-32.

[18] Pollak,M. and Siegmund,D. (1985): A Diffusion Process and its applications to detecting a change in the drift of Brownian Motion. *Biometrika* **72** p267-280

[19] Rogers,L.G.C. and Williams,D. (1987): Diffusions, Markov Processes and Martingales volume 2: Ito Calculus. *Wiley.*

[20] Sprott,D.A. (1965): Transformations and Sufficiency. *Journal of the Royal Statistical Society Series B* **27** p479-485.

[21] Tong,Y.L. (1980): Probability Inequalities in Multivariate Distributions. *Academic Press.*

[22] Wedderburn,R.W.M. (1974): Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** p439-447

[23] Wilkinson,G.N. (1977): On Resolving the controversy in statistical inference. *Journal of the Royal Statistical Society Series B* **39** p119-144.